OPEN ACCESS

BIO-complexity

Critical Review

The Case Against a Darwinian Origin of Protein Folds

Douglas D. Axe*

Biologic Institute, Redmond, Washington, USA

Abstract

Four decades ago, several scientists suggested that the impossibility of any evolutionary process sampling anything but a miniscule fraction of the possible protein sequences posed a problem for the evolution of new proteins. This potential problem—the *sampling problem*—was largely ignored, in part because those who raised it had to rely on guesswork to fill some key gaps in their understanding of proteins. The huge advances since that time call for a careful reassessment of the issue they raised. Focusing specifically on the origin of new protein folds, I argue here that the sampling problem remains. The difficulty stems from the fact that new protein functions, when analyzed at the level of new beneficial phenotypes, typically require multiple new protein folds, which in turn require long stretches of new protein sequence. Two conceivable ways for this not to pose an insurmountable barrier to Darwinian searches exist. One is that protein function might generally be largely indifferent to protein sequence. The other is that relatively simple manipulations of existing genes, such as shuffling of genetic modules, might be able to produce the necessary new folds. I argue that these ideas now stand at odds both with known principles of protein structure and with direct experimental evidence. If this is correct, the sampling problem is here to stay, and we should be looking well outside the Darwinian framework for an adequate explanation of fold origins.

Cite as: 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

Editor: Matti Leisola

Received: December 16, 2009; **Accepted:** March 23, 2010; **Published:** April 15, 2010

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2010.1.c. An early version of this paper (pre-peer review) is scheduled for publication in: Gordon BL, Dembski WA, eds. (2010) The Nature of Nature: Examining the Role of Naturalism in Science. ISI Books (Wilmington).

* Email: daxe@biologicinstitute.org

INTRODUCTION

The elucidation of the genetic code in the late 1960s provided a precise framework for understanding the effects of genetic mutations on protein sequences. Because proteins perform most of the molecular tasks needed for life, solving the code also opened the possibility of understanding the connection between genotype and phenotype on a scale that was not previously possible—the fine scale of nucleotide bases rather than the coarse scale of whole genes. Among other benefits, this promised unprecedented insight into the inner workings of the evolutionary process at the molecular level.

Along with this benefit, however, came a challenging puzzle. The code had made it clear that the vast set of *possible* proteins that could conceivably be constructed by genetic mutations is far too large to have actually been sampled to any significant extent in the history of life. Yet how could the highly incomplete sampling that has occurred have been so successful? How could it have located the impressive array of protein functions required for life in all its forms, or the comparably impressive array of protein structures that perform those functions? This concern was raised repeatedly in the early days of the genetic code [1–4], but it received little attention from the biological community.

One possible reason for the lack of serious attention was the simplicity of the analyses being offered and their reliance on guesswork to fill in for missing data. With fewer than a dozen protein structures deposited in the Protein Data Bank at its founding in 1971 [5, 6], no one at the time had any idea how diverse the complete structural repertoire of biological proteins might be, or what a simple proteome might look like in terms of structural diversity. The functional diversity of proteins was becoming increasingly clear through steady progress in mapping out life's metabolic pathways [7–9], but the all-important questions of how protein structures are suited to functional roles and how protein sequences produce those suitable structures were only just being framed.

With the advantage of a now expansive catalogue of information on whole genomes and proteomes, I aim here to re-examine the puzzle that presented itself four decades ago. I will focus on the origin of substantially new classes of protein structure, or *folds* as they are known. It could be argued that new protein *functions* should instead be the focus because of the direct connection between function and phenotype. But since the primary objective in the origins field is to explain what exists, the great variety of extant protein folds poses an important challenge in its own right. Furthermore, because many new functions seem to have required new folds, explaining these new folds is really one part of the broader challenge of explaining new functions.

From that perspective, the origin of protein folds can be framed with complete generality as a search problem. Briefly, because genes encode proteins, any functional problem that can be solved with a suitable protein can be solved with a suitable gene. Therefore any functional challenge that calls for structural innovation may be thought of as posing a search problem where the search space is the set of possible gene sequences and the target is the subset of genes within that space that are suitable for meeting the challenge. Wherever we see task-specific protein structures in biology, we know that the corresponding search problem was solved successfully—*somehow*. The aim here will be to decide whether Darwinian mechanisms (broadly construed) can reasonably be credited with this success. We will tackle this in two stages. First we will use current knowledge of biological proteins to assess the

difficulty of the search problems that have been solved in simple life forms, and then by assessing the capability of Darwinian searches, we will evaluate the adequacy of the standard model.

ANALYSIS

The problem of sparse sampling

Proteins are natural polymers—large molecules made by connecting smaller building blocks to form unbranched chains. In the general terminology of polymer chemistry, the building blocks are called *monomers*. Amino acids, which come in twenty different kinds, are the monomers used to construct biological proteins. The twenty amino acids differ not in the way they connect to form the main chain, but in their chemically distinct appendages, called *side chains*, that protrude from the main chain. Since there are n^{ℓ} possible ways to construct a polymer chain of length ℓ from n distinct monomer types, amino acid chains a mere twelve residues long (*residue* being the term for an amino acid monomer that has been incorporated into a protein polymer) can be built in four quadrillion ways ($20^{12} = 4 \times 10^{15}$). The gene sequences that encode these short chains are even more numerous as a consequence of the many-to-one mapping of *codons* (the nucleotide triplets of the genetic code) to encoded amino acids.

For either kind of sequence, gene or protein, the number of distinct sequence possibilities grows very rapidly as the chain length is increased. Focusing principally on proteins, we begin by asking how long these biological polymers tend to be. The answer to this will tell us how large the relevant sequence space is. It should be emphasized, though, that this is only a starting point. Several other aspects of proteins will need to be examined before we can decide whether their size complicates Darwinian explanations of their origins.

The simple relationship between gene sequences and protein sequences in bacteria allows protein sizes to be determined directly from genomic data. This, in combination with abundant data on protein structures and functions, makes the well studied gut bacterium *Escherichia coli* an excellent model system for examining a simple proteome.[1] The size of *E. coli* proteins with known functions can be assessed by analyzing the data files provided by EcoCyc [10], a comprehensive database for this organism. Figure 1 shows the distribution of protein chain lengths for all proteins known to be involved in enzymatic functions in *E. coli*, either alone or in combination with other proteins. From the mode of the distribution we see that the most common length of these proteins is around 300 amino acid residues, with the higher mean and median lengths reflecting the existence of numerous protein chains that are much longer than this.

If we take 300 residues as a typical chain length for functional proteins, then the corresponding set of amino acid sequence possibilities is unimaginably large, having 20^{300} (= 10^{390}) members. How or whether this number should figure into our assessment of origins scenarios will be examined in the following sections. Here the point is simply that biological protein sequences are indeed members of astoundingly large sets of sequence possibilities. And by 'astoundingly large' we mean much more numerous than any mutation events we might postulate as having produced them. According to one estimate, the maximum number of distinct physical events that could have occurred within the visible universe, including all particles throughout the time since the Big Bang, is 10^{150} [11]. Since only a minute fraction of these events had anything to do with producing new protein sequences, we can assert with confidence that there is a vast disparity between the number of distinct protein sequences of normal length that are *possible*, on the one hand, and the number that might have become *actual*, on the other. In other words, real events have provided only an exceedingly sparse sampling of the whole set of sequence possibilities.

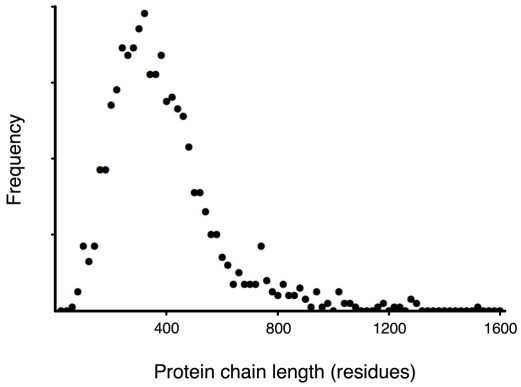

Figure 1. Approximate length distribution for 1,326 proteins known to be enzymes or enzyme components in *E. coli*. The mean and median lengths are 389 residues and 350 residues, respectively. Molecular weights of protein chains, obtained from EcoCyc version 9.0 (*proteins.dat* data file), were converted to approximate chain lengths by using an average per-residue molecular weight of 110 g/mol. Lengths were binned in 20-residue increments, the most occupied bin containing 78 protein chains. **doi:**10.5048/BIO-C.2010.1.f1

We will refer to this as the problem of *sparse sampling*, or the *sampling problem*, with the intent of deciding whether or not it really is a problem for the standard evolutionary model. At the very least it raises the important question of how such sparse sampling would uncover so many highly functional protein sequences. To picture the difficulty, imagine being informed that a valuable gemstone was lost somewhere in the Sahara Desert. Without more specific information, any proposal for finding the missing gem would have to come to terms with the vastness of this desert. If only an infinitesimal fraction of the expanse can feasibly be searched, we would judge the odds of success to be infinitesimally small.

Evolutionary searches for functional proteins might seem less hopeless in some respects, though. For one, there is a highly many-to-one mapping of protein sequences onto protein functions. This means that vast numbers of comparably valuable targets (protein sequences that are comparably suitable for any particular function) are there to be found. Therefore, while it is effectively impossible to stumble upon a particular 1-in-10^{390} protein sequence by chance, the likelihood of stumbling upon a particular protein *function* by chance will be m-fold higher, where m represents the multiplicity of sequences capable of performing that function.

There are good reasons to be cautious about this, however. Natural proteins would have to be much larger than they need to be and/or highly indifferent to the specifics of their amino acid sequences in order for m to be large enough to resolve the problem. We can imagine a different world where, for example, the planetary surface has rich deposits of abiotic amino acids, and cells indiscriminately incorporate these amino acids into long polypeptide chains, and these chains somehow benefit the cells without performing complex functions. In that world the problem we address here would not exist. But in our world things are strikingly different. Here we see a planet with amino acids of strictly biological origin, and we see cells going to extraordinary lengths to manufacture, use, recycle, and scavenge all twenty of them. We see elaborate error-checking mechanisms that minimize the

[1] The term *proteome* refers to the complete set of proteins in an organism or cell type.

chances of confusing any one amino acid for any other during protein synthesis, and (as already noted) we see that the products of this tightly controlled process are *long* proteins. Lastly, we see that these long proteins perform an impressive variety of functions with equally impressive specificity and efficiency.

In the face of this, either we accept that proteins are what they seem to be—long amino acid chains that need to meet stringent sequence requirements in order to work—or we suppose that they are really much simpler than they seem to be. It has to be said that the second option arouses immediate suspicion by disregarding matters of plain fact—both the actual properties of proteins and the cellular processes that only make sense if the first option is correct. This is admittedly more of a suggestion than a proof, and yet it does clearly add to the burden of justifying the second option. More conclusive arguments will require a closer look at the data.

The need for large proteins

Enzymes are proteins or protein complexes that perform chemical transformations in a highly efficient and specific way. When students first encounter them, one of the things they may find puzzling is that they tend to be quite large in comparison to their active sites—the part that actually binds the reactants (or *substrates*, as they are known) and converts them into products. Catalase, for example, is an enzyme consisting of four identical protein chains with individual molecular weights of around 80,000 g/mol. Each chain forms an active site that functions as an extremely efficient converter of individual hydrogen peroxide (H_2O_2) molecules into water and oxygen. At 34 g/mol, though, this tiny substrate molecule has less than 1/2000th the mass of the protein that works on it. Mass ratios differ widely from one enzyme to the next, but as a rule small-molecule metabolism employs enzymes that are very large in comparison to their substrates.

Why are these enzymes so much larger than the things they manipulate? Although we are some way from a complete answer to this, several aspects of the relationship between enzyme structures and their functions provide at least partial answers. On the most basic level, it has become clear that protein chains have to be of a certain length in order to fold into stable three-dimensional structures. This requires several dozen amino acid residues in the simplest structures, with more complex structures requiring much longer chains. In addition to this minimal requirement of stability, most folded protein chains perform their functions in physical association with other folded chains [12]. The complexes formed by these associations may have symmetrical structures made by combining identical proteins or asymmetrical ones made by combining different proteins. In either case the associations involve specific inter-protein contacts with extensive interfaces[2]. The need to stabilize these contacts between proteins therefore adds to their size, over and above the need to stabilize the structures of the individual folded chains.

Beyond these general principles, we have specific understanding of the need for structural complexity with respect to many particular protein functions. In catalase, for example, the active sites are deeply buried within the enzyme, such that the H_2O_2 molecules must pass through a long channel before they are catalytically converted. By replacing some of the amino acids in the enzyme, it has been shown that an electrical potential gradient within the channel makes an important contribution to the catalytic process [13]. So in this case, as in many others, the enzyme has important interactions with the substrate some distance away from the place where the actual chemical conversion occurs. We see in such examples that enzymes may guide reactants and/or products through a process that is more extensive than mere catalysis, and this processing requires a structure than extends well beyond the active site.

Another common functional constraint with implications for protein size is the need to achieve direct coupling of processes occurring at different sites on the same enzyme. This is distinct from the indirect coupling commonly achieved by diffusion of shared metabolites within the cell. Direct coupling, unlike simple diffusion, has to be mediated by structural connections between the sites being coupled, and this requires more extensive protein structures.

Three examples will illustrate the importance of coupling in biology. The first is carbamoyl phosphate synthetase (CPS), which has been aptly described as "a remarkably complex enzyme" [14]. It uses bicarbonate, glutamine, and ATP to make carbamoyl phosphate, which is required for the biosynthesis of both arginine and pyrimidine ribonucleotides. In order to couple the reactions occurring at its three active sites, this enzyme uses internal molecular tunnels for efficient transfer of reactants. To achieve this channel-coupled multi-site architecture, CPS uses two protein chains with a combined length of over 1,400 amino acid residues (Figure 2). Thoden and co-workers describe the design rationale for this complexity as follows:

> From extensive biochemical data, it is now known that a fully coupled CPS requires the hydrolysis of one glutamine and the utilization of two molecules of MgATP for every molecule of carbamoyl phosphate formed. The three active sites of the enzyme must therefore be synchronized with one another in order to maintain the overall stoichiometry of the reaction without wasteful hydrolysis of glutamine and/or MgATP [14].

So again we see an enzyme having to coordinate a complex process rather than a simple reaction, and having to be large in order to achieve this.

As a second example of direct coupling, consider the following representations of cellular processes:

$$H^+_{ext} \rightarrow H^+_{int}\text{, and}$$
$$ADP + P_i \rightarrow ATP + H_2O.$$

The first describes the flow of protons down an electrochemical potential gradient from the exterior of a membrane-enclosed compartment to the interior, and the second describes the generation of ATP from ADP and inorganic phosphate. These two processes have no essential connection. That is, there is no general principle of physics or chemistry by which ATP synthesis and proton fluxes have anything to do with each other. From an engineering perspective, however, it is often possible and desirable to design devices that *force* a relation upon otherwise unrelated processes. Of particular interest in this regard are devices like solar cells and turbines that harness energy from an available source in order to accomplish useful tasks that require energy.

Life likewise crucially depends on many such devices, one of which provides highly efficient energetic coupling of the above two processes. This coupler, the proton-translocating ATP synthase, is a rotary engine built from eight or more protein types, some of which are used multiple times to form symmetric substructures (Figure 3). Various versions of this ingenious device are found in all forms of life. The mitochondrial version couples the processes in the direction shown above, with an energetically favorable proton flux driving the energetically unfavorable (but biologically crucial) synthesis of ATP. The mechanism by which it operates is fascinating, but for the present purposes the key

[2] The median interfacial area of protein-protein interfaces in the ProtCom database is 975 Å^2 (http://www.ces.clemson.edu/compbio/protcom/Search40.htm).

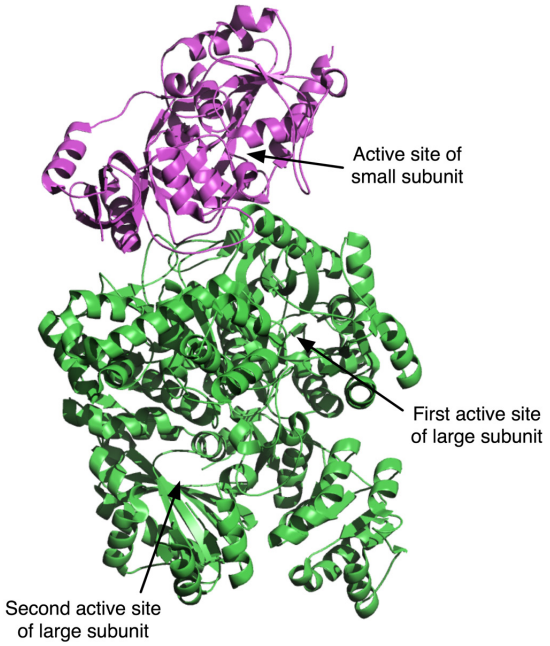

Figure 2. Structure of carbamoyl phosphate synthetase from *E. coli*. In this common way of representing protein structure, alpha helices are shown as coiled ribbons and beta strands are shown as ribbons with arrowheads. The two proteins chains (differentiated by color) are rendered according to Protein Data Bank (PDB [6]) entry 1T36.
doi:10.5048/BIO-C.2010.1.f2

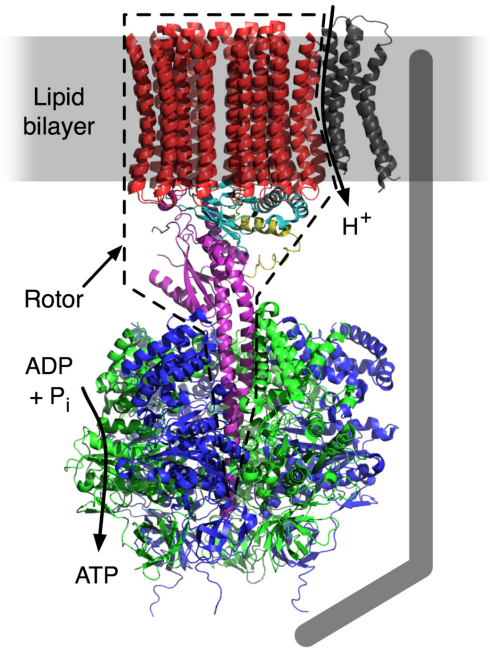

Figure 3. Partial structure of the proton translocating ATP synthase. The energy liberated by proton flow causes rapid rotation of the rotor, which is an assembly of many protein chains (colors indicating chain identity). The other protein chains form the stator (the stationary part of the motor). The bent grey bar shows the approximate location of a portion of the stator for which the structure is not fully known. The proteins used to construct the *E. coli* synthase (some present in multiple copies) have 79, 139, 156, 177, 271, 287, 460, and 513 amino acid residues, for a combined length (non-redundant) of 2,082 residues. The representation is based on PDB entries 1C17 and 1E79.
doi:10.5048/BIO-C.2010.1.f3

point is that the crucial function it performs absolutely requires an overall structure that is very large, which presents a very large sampling problem. The ATP synthase has to manage the flow of protons through the lipid bilayer in such a way as to produce the rotation of a molecular rotor, which in turn forces a conformational shift in the portion of the stator that contacts the other end of the rotor, which in turn causes cyclic binding and phosphorylation of ADP, followed by release of the desired product—ATP. Clearly only a substantial protein structure could possibly orchestrate a process of this physical and spatial complexity. Indeed, in view of what it accomplishes, what amazes us is how *small* it is.

The ATP synthase provides an opportunity at this point to refine the connection between protein size and the sampling problem. Returning to the lost gemstone metaphor, the gem is a new beneficial function that can be provided by a protein or a set of proteins working together, and the desert is the whole space of sequence possibilities within which successful solutions are to be found. Although some of the component proteins that form the ATP synthase are at the small end of the distribution shown in Figure 1 (see Figure 3 legend), none of these performs a useful function in itself. Rather, the function of ATP production requires the whole suite of protein components acting in a properly assembled complex. Consequently, the desert is most precisely thought of as the space of all DNA sequences long enough to encode that full suite. For our purposes, though, it will suffice to picture the space of protein sequences of a length equaling the combined length of the different protein types used to form the working complex (around 2,000 residues for the ATP synthase; see Figure 3 legend). This takes into account both the need for multiple non-identical chains in many working structures and the sequence redundancy that exists when multiple identical chains are used. It also dramatically expands the size of the search space in the common case where a protein chain of one kind is useful only in combination with other kinds.

As a final example of the role of coupling in biology, we return to the connection between protein sequences and the DNA sequences that encode them. When expressed as an abstract mapping of codons to amino acids, this is the familiar genetic code often represented in table form. The physical embodiment of that code, though, is the set of aminoacyl-tRNAs—large RNA derivatives incorporating both the anticodon loops that 'recognize' their respective codons on mRNA and the amino acids that these codons specify. Indeed, the genetic code only has its law-like status because aminoacyl-tRNAs are synthesized with anticodons paired very reliably with their corresponding amino acids. Because the anticodon loops in tRNAs are spatially distant from the amino acid attachment sites (Figure 4), the enzymes that accomplish this reliable pairing have to be large in order to attach the amino acids while simultaneously 'verifying' that they are the correct ones.

Many more examples could be given (e.g., [15]) but the ones we have examined adequately make the point that cellular functions often require large proteins, which leads to a large sampling problem. When we consider the sets of distinct proteins that commonly provide these functions, the sampling problem becomes even more challenging. Extreme examples abound. Ribosomes, for example, depend on the coordinated action of some fifty different proteins in order to synthesize new proteins. But even functions of more typical complexity amply demonstrate that the challenge of sparse sampling goes all the way back to the origin of protein-catalyzed metabolism and genetic processing. The many functions involved in gene expression had to be in place from the outset, and because these functions require large protein structures, this means the sampling problem appeared as soon as the genetic code appeared.

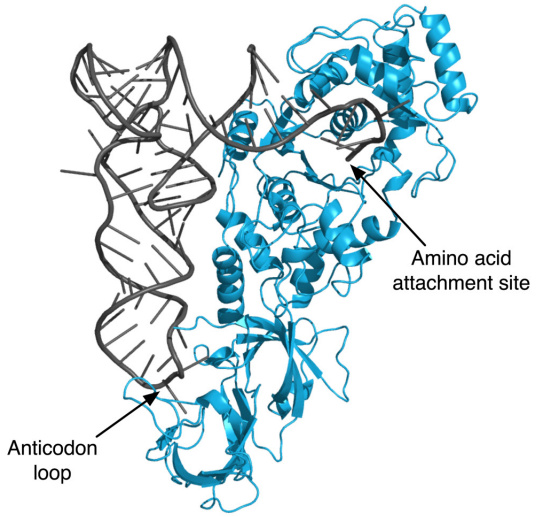

Figure 4. Structure of *E. coli* glutaminyl-tRNA synthetase. RNA bases are here rendered as sticks protruding from the RNA backbone; the enzyme is a single protein chain of 554 amino acid residues (both rendered according to PDB entry 1GTR). **doi:**10.5048/BIO-C.2010.1.f4

The rarity of functional folds

Having shown that the problem of sparse sampling is *real*—meaning that cellular functions require proteins or suites of proteins that are of necessity far too large for the sequence possibilities to have been sampled appreciably—we now turn to the question of whether it is really a *problem* for neo-Darwinian evolution. Two possibilities for mitigating the problem need to be considered. One of these has been mentioned already. It is the possibility that the multiplicity of sequences capable of performing the requisite functions, *m*, might be large enough for working sequences to be found by random searches. The second possibility is that functional protein sequences might bear a relationship to one another that greatly facilitates the search. In the desert metaphor, imagine all the different gems being together in close proximity or perhaps lined up along lines of longitude and latitude. In either of these situations, or in others like them, finding the first gem would greatly facilitate finding the others because of the relationship their positions bear to one another.

We will complete our examination of the first of these possibilities before moving to the second. As noted previously, for *m* to be large enough to compensate for the size of the search spaces, modern proteins would have to be either much larger than necessary or constructed with much higher sequence fidelity than necessary. I pointed out above that both of these assumptions, superfluous size and superfluous sequence fidelity, make it hard to explain the sophisticated cellular systems that are devoted to producing the twenty amino acids and precisely incorporating them into long protein chains. Subsequently I presented the evidential case for rejecting the notion of superfluous size. What remains is to consider the specific evidence pertaining to the possibility of superfluous fidelity.

Superfluous fidelity implies that protein synthesis is much more fastidious about amino acid identities than protein function is. Consequently, we can reframe this possibility in terms of functional constraints. Namely, for *m* to be large enough to resolve the sampling problem, it would have to be the case that protein functions place very loose constraints on amino acid sequences. Although many studies have sought to assess these constraints, the conclusions drawn seem inconsistent, some studies describing the constraints as very loose (e.g., [16–20]), while others find them to be very stringent (e.g., [21–25]).

To untangle the apparent contradiction, we need to quantify a *boundary* value for *m*, meaning a value which, if exceeded, would solve the whole sampling problem. To get this we begin by estimating the maximum number of opportunities for spontaneous mutations to produce any new species-wide trait, meaning a trait that is fixed within the population through natural selection (i.e., selective sweep). Bacterial species are most conducive to this because of their large effective population sizes.[3] So let us assume, generously, that an ancient bacterial species sustained an effective population size of 10^{10} individuals [26] while passing through 10^4 generations per year. After five billion years, such a species would produce a total of 5×10^{23} ($=5 \times 10^9 \cdot 10^4 \cdot 10^{10}$) cells that happen (by chance) to avoid the small-scale extinction events that kill most cells irrespective of fitness. These 5×10^{23} 'lucky survivors' are the cells available for spontaneous mutations to accomplish whatever will be accomplished in the species. This number, then, sets the maximum probabilistic resources that can be expended on a single adaptive step. Or, to put this another way, any adaptive step that is unlikely to appear spontaneously in that number of cells is unlikely to have evolved in the entire history of the species.

In real bacterial populations, spontaneous mutations occur in only a small fraction of the lucky survivors (roughly one in 300 [27]). As a generous upper limit, we will assume that *all* lucky survivors happen to receive mutations in portions of the genome that are not constrained by existing functions[4], making them free to evolve new ones. At most, then, the number of different viable genotypes that could appear within the lucky survivors is equal to their number, which is 5×10^{23}. And again, since many of the genotype differences would not cause distinctly new proteins to be produced, this serves as an upper bound on the number of new protein sequences that a bacterial species may have sampled in search of an adaptive new protein structure.

Let us suppose for a moment, then, that protein sequences that produce new functions by means of new folds are common enough for success to be likely within that number of sampled sequences. Taking a new 300-residue structure as a basis for calculation (I show this to be modest below), we are effectively supposing that the multiplicity factor *m* introduced in the previous section can be as large as $20^{300}/ 5{\times}10^{23} \approx 10^{366}$. In other words, we are supposing that particular functions requiring a 300-residue structure are realizable through something like 10^{366} distinct amino acid sequences. If that were so, what degree of sequence degeneracy would be implied? More specifically, if 1 in $5{\times}10^{23}$ full-length sequences are supposed capable of performing the function in question, then what proportion of the twenty amino acids would have to be suitable on average at any given position? The answer is calculated as the 300^{th} root of $(5{\times}10^{23})^{-1}$, which amounts to about 83%, or 17 of the 20 amino acids. That is, by the current assumption proteins would have to provide the function in question by merely *avoiding* three or so unacceptable amino acids at each position along their lengths.

No study of real protein functions suggests anything like this degree of indifference to sequence. In evaluating this, keep in mind that the indifference referred to here would have to characterize the *whole* protein rather than a small fraction of it. Natural proteins commonly tolerate some sequence change without com-

[3] Stochastic aspects of survival having nothing to do with fitness make it very unlikely that any particular instance of beneficial mutation will lead to fixation of the resulting genotype. Roughly speaking, the effective size of a real population is the size of a hypothetical population lacking these stochastic influences that is as conducive to fixation of new genotypes as the real one is.

[4] This presupposes a much higher tolerance of non-functional ('junk') DNA than modern bacteria exhibit.

plete loss of function, with some sites showing more substitutional freedom than others. But this does not imply that most mutations are harmless. Rather, it merely implies that complete inactivation with a single amino acid substitution is atypical when the starting point is a highly functional wild-type sequence (e.g., 5% of single substitutions were completely inactivating in one study [28]). This is readily explained by the capacity of well-formed structures to sustain moderate damage without complete loss of function (a phenomenon that has been termed the *buffering effect* [25]). Conditional tolerance of that kind does not extend to whole proteins, though, for the simple reason that there are strict limits to the amount of damage that can be sustained.

A study of the cumulative effects of conservative amino acid substitutions, where the replaced amino acids are chemically similar to their replacements, has demonstrated this [23]. Two unrelated bacterial enzymes, a ribonuclease and a beta-lactamase, were both found to suffer complete loss of function *in vivo* at or near the point of 10% substitution, despite the conservative nature of the changes. Since most substitutions would be more disruptive than these conservative ones, it is clear that these protein functions place much more stringent demands on amino acid sequences than the above supposition requires.

Two experimental studies provide reliable data for estimating the proportion of protein sequences that perform specified functions. One study focused on the AroQ-type chorismate mutase, which is formed by the symmetrical association of two identical 93-residue chains [24]. These relatively small chains form a very simple folded structure (Figure 5A). The other study examined a 153-residue section of a 263-residue beta-lactamase [25]. That section forms a compact structural component known as a *domain* within the folded structure of the whole beta-lactamase (Figure 5B). Compared to the chorismate mutase, this beta-lactamase domain has both larger size and a more complex fold structure.

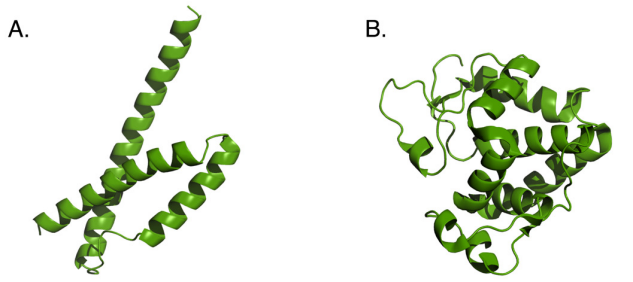

Figure 5. Structures of protein chains used to measure the rarity of working sequences. A) A single chain (93 residues) from the AroQ-type chorismate mutase examined by Taylor *et al.* [24] (PDB entry 1ECM). B) The 153-residue portion of the TEM-1 beta-lactamase examined by Axe [25] (PDB entry 1ERM). **doi:**10.5048/BIO-C.2010.1.f5

In both studies, large sets of extensively mutated genes were produced and tested. By placing suitable restrictions on the allowed mutations and counting the proportion of working genes that result, it was possible to estimate the expected prevalence of working sequences for the hypothetical case where those restrictions are lifted. In that way, prevalence values far too low to be measured directly were estimated with reasonable confidence.

The results allow the average fraction of sampled amino acid substitutions that are functionally acceptable at a single amino acid position to be calculated. By raising this fraction to the power ℓ, it is possible to estimate the overall fraction of working sequences expected when ℓ positions are simultaneously substituted (see reference 25 for details). Applying this approach to the data from the chorismate mutase and the beta-lactamase experiments gives a range of values (bracketed by the two cases) for the prevalence of protein sequences that perform a specified function. The reported range [25] is one in 10^{77} (based on data from the more complex beta-lactamase fold; $\ell = 153$) to one in 10^{53} (based on the data from the simpler chorismate mutase fold, adjusted to the same length: $\ell = 153$). As remarkable as these figures are, particularly when interpreted as probabilities, they were not without precedent when reported [21, 22]. Rather, they strengthened an existing case for thinking that even very simple protein folds can place very severe constraints on sequence.

Rescaling the figures to reflect a more typical chain length of 300 residues gives a prevalence range of one in 10^{151} to one in 10^{104}. On the one hand, this range confirms the very highly many-to-one mapping of sequences to functions. The corresponding range of m values is 10^{239} ($=20^{300}/10^{151}$) to 10^{286} ($=20^{300}/10^{104}$), meaning that vast numbers of viable sequence possibilities exist for each protein function. But on the other hand it appears that these functional sequences are nowhere near as common as they would have to be in order for the sampling problem to be dismissed. The shortfall is itself a staggering figure—some 80 to 127 orders of magnitude (comparing the above prevalence range to the cutoff value of 1 in 5×10^{23}). So it appears that even when m is taken into account, protein sequences that perform particular functions are far too rare to be found by random sampling.

Shortcuts to new folds?

The possibility yet to be examined is that functional protein sequences might bear a relationship to one another that allows spontaneous mutations to discover new functional protein folds much more readily than wholly random sampling would. The simplest way for this to occur would be if *all* functional sequences, regardless of what their functions are, happen to be much more similar to each other than a pair of random sequences would be. In other words, suppose there were a universal consensus sequence that typified all biological proteins, with functional diversity caused by minor deviations from that consensus. The effect of such a universal correlation between sequence and function would be to concentrate all the useful protein sequences within a tiny region of sequence space, making searches that start in that region much more likely to succeed.

Localized searches of this kind are known to work in some cases. Many enzymes, for example, can be made to perform their catalytic functions on different substrates by changing just one or two amino acids within their active sites. Bacteria often evolve resistance to modified versions of existing antibiotics in this way, by an existing resistance enzyme acquiring expanded substrate specificity [29]. The evolutionary search for resistance to the new antibiotic works in these cases because the original enzyme needs only slight adjustment in order to perform the new task. Consequently a local search of point-mutation variants has a reasonably good chance of succeeding.

The problem comes when we attempt to generalize this local phenomenon. Although there are definite correlations between the various kinds of functions that proteins perform and the respective fold structures used to perform them, and these structural correlations often imply sequence correlations as well, it is simply not the case that *all* functional folds or sequences are substantially alike. Consequently, while local searches may explain certain local functional transitions, we are left with the bigger problem of explaining how so many fundamentally new protein structures and functions first appeared.

To get an idea of the scale of this problem, consider that the SCOP classification of protein structures currently has 1,777 different structural categories for protein domains, the basic units

of folded protein structure.[5] But since that count only includes known protein structures, it is certainly an underestimate. Each new genome project reveals numerous protein sequences with no significant similarity to any previously known sequence, which suggests that the actual number of fundamentally distinct protein domains may be much higher [32]. Whatever the true figure turns out to be, it is clearly large enough that no model of protein origins can be considered satisfactory without accounting for the origin of this great variety of domain folds.

In fact, although the sampling problem has here been framed in terms of protein chains, it could equally be framed in terms of domains. Since domains are presumed to be the fundamental units of conserved structure in protein evolution [33], the question of whether functional sequences are confined to a small patch of sequence space is best addressed at the domain level. And it turns out that domain sequences are not confined in this way. When structurally unrelated protein domain sequences are aligned optimally, the resulting alignment scores are very similar to the expected scores for randomized sequences with the same amino acid composition [34][6]. Since random sequences produced in this way are widely scattered through sequence space, this means that dissimilar natural sequences are as well. In fact, because amino acid composition correlates with structural class [35], we would expect random sequences with average compositions to align somewhat better than dissimilar natural sequences do. Indeed, such wide dispersion of natural domain sequences throughout sequence space is not surprising considering the great variety of domain structures that these sequences form (Figure 6).

However, the broad parallel between structural diversity and sequence diversity in modern proteins might possibly be explained by structural and functional divergence having occurred first, with sequence divergence following later. If transitions to new folds can be expected to occur with minimal sequence alteration, then perhaps the wide dispersal of modern proteins throughout sequence space is a *consequence* of structural divergence rather than a *precondition* of it. A recent study by Alexander and co-workers [36] may seem to support this. By showing that a single amino acid residue can act as a conformational switch, causing a small protein to adopt one or the other of two substantially different domain structures depending on which of two amino acids (leucine or tyrosine) is present at the key position, they have demonstrated the surprising possibility of minimal sequence alteration causing a significant structural transition.

Their further demonstration of a functional transition makes this study particularly relevant to our present discussion. The sequence context in which the conformational switch works was constructed by making two dissimilar natural sequences (which produce the two dissimilar structures) progressively more similar until a single amino acid difference became the deciding factor in stabilizing one fold over the other. Since the two structures have different functions, and a functional transition was shown to accompany the structural transition, the authors have demonstrated a stepwise mutational pathway between different folds in which "neither function nor native structure is completely lost" [36].

As always, though, discerning the implications for natural evolution calls for caution. In particular, proteins treated as functional for the purposes of a laboratory experiment may not meet the more complex demands of biological function. The two domains in this study come from a cell wall protein, called protein G, that certain pathogenic bacteria use to evade immune detection. By latching on to particular serum proteins in the blood of an infected host, these binding domains enable an invading bacterium to "camouflage" itself with host proteins, so as not to trigger an immune response [37]. The modified domains constructed by Alexander and co-workers do bind the serum proteins, but under mush less challenging conditions. In particular, rather than having to discriminate the appropriate serum proteins from all the other proteins present in blood, the modified domains merely had to show an affinity for their binding partners as purified components, and moreover at a reduced temperature (which favors the desired association) [36].

Since function enters the evolutionary process strictly in terms of fitness, this disparity of conditions presents a problem. The study's authors infer from their findings that "nature will explore sequence space when there is no penalty for doing so—that is, nature will follow any functional path" [36]. However their demonstrated path to new structure and function *does* appear to involve a significant fitness penalty. The natural protein G domains are stably folded at the temperature of human blood (37 °C), which is presumably a requirement for their biological function. Yet neither of the conformation-switch variants are stable at this temperature [36]. Consistent with this, the variants are found to be less proficient at binding the purified serum proteins than the natural domains are, even under the favorable laboratory conditions [36]. So, while it is true that neither structure nor function is completely lost along the mutational path connecting the two natural sequences, natural selection imposes a more stringent condition. It does not allow a population to take *any* functional path, but rather only those paths that carry no fitness penalty.

The vastness of sequence space imposes another restriction. Based on the considerations of the previous section, we should expect that the protein G domain structures are specified by very minute proportions of the possible amino acid sequences of similar length. Their very small size (roughly fifty amino acids) means these proportions should be orders of magnitude greater than for larger folds, but minute nonetheless.[7] The fascinating finding of Alexander and co-workers is that these two distinct subsets[8] of protein sequence space have what appears to be a single point of contact—the switch point where they differ by only one residue. But based on the reported protein sequences [36, 38], the shortest mutational path from one of the natural domain sequences to the alternative structure consists of 21 amino acid substitutions that require 30 nucleotide changes. Consider, then, that of all the possible ways to mutate the natural starting sequence to this extent, it appears that only *one* of them produces the new fold. From the vantage point of that starting sequence, this makes successful fold conversion an exceedingly remote possibility in terms of sheer mutational odds—one sequence among 10^{46} equally accessible alternatives.[9] Combined with the apparent fitness penalty, we conclude that the demonstrated transition, while clearly interesting in other respects, does not solve the problem of sparse sampling.

In the end, the evolutionary difficulty that Alexander and coworkers point to in their introduction—that mutational paths to completely new folds will inevitably destabilize the original fold before producing the new one [36]—remains valid in light of their

[5] SCOP version 1.73 [30] organizes domain structures into 1,777 categories at the 'superfamily' level, based on "structures and, in many cases, functional features [that] suggest a common evolutionary origin." [31]

[6] The *Z*-score of an alignment compares the raw alignment score to the raw scores of optimally aligned randomized versions of the initial pair of sequences. The *Z*-scores plotted in reference 34 for dissimilar domain sequences are distributed around zero, meaning that the actual alignments tend to be comparable to randomized alignments.

[7] For the 153-residue beta-lactamase domain (Figure 5B) the proportion was estimated as 1 in 10^{97} [25]. A protein one third this size with similar per-residue constraints would be specified by roughly 1 sequence in 10^{32}.

[8] By definition a folded structure is only stable if it is the predominant conformation under specified conditions. Consequently sets of sequences specifying two different stable folds cannot overlap.

[9] The number of ways to change 30 nucleotide bases in a stretch of 150 bases is 3^{30} $(150!/(120!\,30!)) = 10^{46}$.

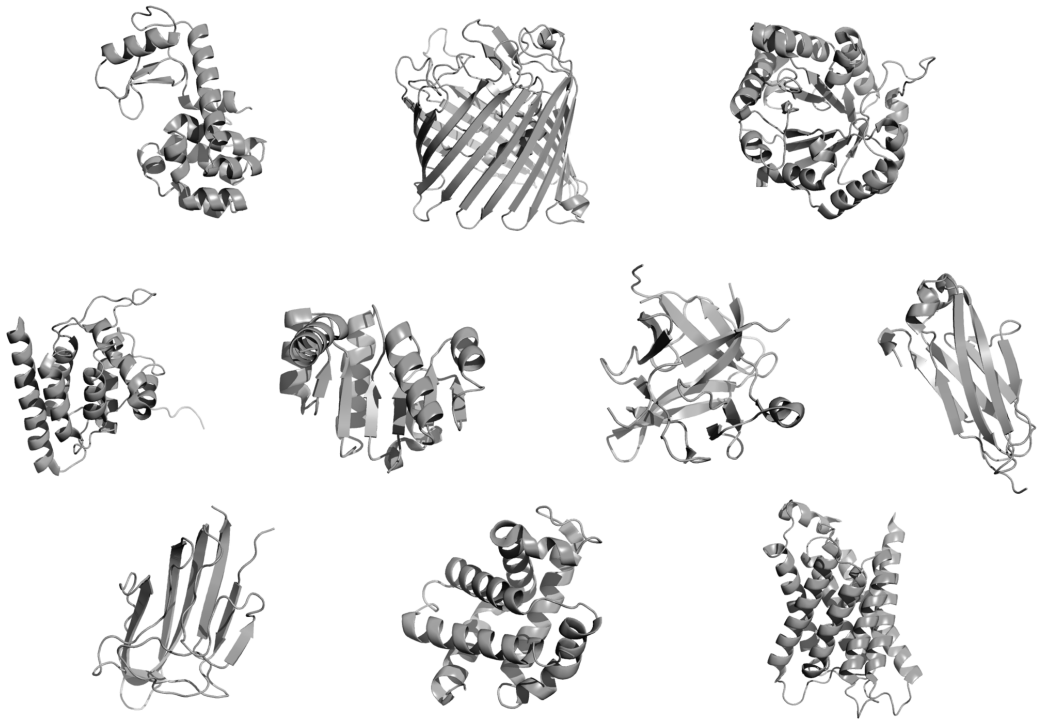

Figure 6. A sample of the structural variety of protein domain folds. Shown in the top row (left to right): bacteriophage T4 lysozyme (PDB: 167L), ompF porin from *E. coli* (PDB: 2OMF), human triosephosphate isomerase (PDB: 1HTI); middle row: run domain of mouse rap2-interacting protein X (PDB: 2CXF), a transport protein from *B. subtilis* (PDB: 1LSU), human interleukin 1-beta (PDB: 1L2H), mouse antibody light chain (PDB: 1UM5); bottom row: a fragment of human collagen (PDB: 1GR3), human hemoglobin alpha chain (PDB: 1IRD), sheep aquaporin (PDB: 1SOR).
doi:10.5048/BIO-C.2010.1.f6

findings. For very small folds like the ones they examined, the buffering effect mentioned in the previous section may allow experimenters to achieve mutational transitions in which none of the intermediates is predominantly unfolded. The reason for this is that well formed small structures can be similar in thermodynamic stability to full-sized structures. As a typical example, aspartate aminotransferase, at 396 residues, has a stability of 10 kilocalories per mole [39], which makes it only slightly more stable than the natural G protein domains. From an experimental standpoint (not in nature), thermodynamic stability is like a bank account that can be drawn down to zero. This enables the stability costs of a carefully engineered transition to a new fold to be borne much more feasibly in the case of small folds, since they have far fewer residues and therefore require fewer changes (21 amino acid changes amount to nearly half of a protein G domain, but only about 5% of aspartate aminotransferase). But real evolution has to happen in nature, where mutations are not carefully chosen and proteins are only as stable as natural selection compels them to be. There are no cash reserves. Withdrawals come at a cost, and selection does not tolerate costs without immediate recompense. So the expectation that substantially different protein structures are best achieved with substantially different protein sequences stands in light of this work, making it fully consistent with what we actually see in natural proteins.

It therefore seems inescapable that considerable distances must be traversed through sequence space in order for new protein folds to be found. Consequently, any shortcut to success, if it exists, must work by traversing those distances more effectively rather than by shortening them. The only obvious possibility here is that new folds might be assembled by recombining sections of existing folds [40–42]. If modular assembly of this kind works, it would explain how just one or two gene fusion events might produce a new protein that differs substantially from its 'parents' in terms of overall sequence and structure. Of course, probabilistic limitations would need to be addressed before this could be deemed a likely explanation (because precise fusion events are much less likely than point mutations), but the first question to ask is whether the assumed modularity is itself plausible.

To examine this further, we begin by considering what this kind of modularity would require. If it is to be of general use for building up new folds, it seems to require that folds be divisible into more or less self-contained structural components that can be recombined in numerous ways, with each combination having a good chance of producing a well-formed composite structure. Two physical criteria would have to be met for this to be true. First, the sequence specificity for forming these components must be internal to the components themselves (making their structures self-contained), and second, the interactions that hold neighboring components together to form composite structures must be *generic* in the sense of lacking critical dependence on the particulars of the components.

The immediate problem is that the first criterion tends to be met only at the level of a complete fold—a folding domain. Important structural features are certainly discernable at lower levels, the most ubiquitous of these being the regular chain conformations known as the alpha helix and the beta strand (*secondary structure* being the term for these repetitive patterns in local chain structure). But these only find stable existence in the context of larger fold structures (*tertiary structure*) that contain them. That is, the smallest unit of protein structure that forms stably and spontaneously is typically a complete globular assembly with multiple, layered elements of secondary structure. Smaller pieces of structure can have some tendency to form on their own, which is important for triggering the overall folding process [43], but the highly cooperative nature of protein folding [44] means that stable structure forms all at once in whole chunks—domains—rather than in small

pieces. Consequently, self-contained structural modules only become a reality at the domain level, which makes them unhelpful for explaining new folds at that level.

The second criterion sheds some light on this. The generic binding it calls for would have to involve a generic aspect of protein structure, which implies an aspect that is sequence-independent. Structurally speaking, protein sequences correspond to the succession of side chains along the main chain, or *backbone* as it is often called. So the backbone, considered not in its folded form but as a flexible molecular chain, is the generic part of a protein. By examining only this generic part, Pauling, Corey and Branson were able to predict the alpha helix and the beta sheet as the two regular recurring geometries in protein chains [45, 46].

Part of what made this prediction possible was the fact that the possibilities are so highly restricted. As Figure 7 shows, both alpha helices and beta sheets (groups of beta strands bound edge-to-edge) primarily present side chains at their exterior. Only the ends of helices and the outer edges of sheets present a generic backbone interface for binding. But the only possible generic additions at these interfaces are those that extend the regular structure, either by elongating the helix or by adding a strand to the edge of the sheet. Although examples of proteins accommodating these structural changes certainly exist [47, 48], the fact that both simply extend existing structure makes them unhelpful for explaining wholesale structural reorganization.

Because structural reorganization requires elements of secondary structure to be grouped spatially in new ways, it necessarily involves new binding interfaces where the exteriors of helices and/or sheets must adhere to each other in new ways. But since these interfaces consist largely of side chains, they are necessarily sequence-dependent and therefore *non*-generic. This is important enough to be restated: *The binding interfaces by which elements of secondary structure combine to become units of tertiary structure are predominantly sequence dependent, and therefore not generic.* This presents a major challenge for the idea of modular assembly of new folds, at least as a general explanation.

The preceding paragraphs develop this challenge in terms of general aspects of protein structure. But it also finds specific experimental support. As we will see next, several studies demonstrate that proteins with substantially different amino acid sequences (roughly 50% amino acid identity or less) fail to show part-for-part structural equivalence even if they are highly similar in terms of overall structure and function. Since the modularity hypothesis assumes a much more demanding sense of structural equivalence (where modules retain their structure even when moved between proteins that differ radically in terms of overall structure and function) the failure of the less demanding sense seems to rule that hypothesis out.

One of these studies used a pair of beta lactamases with indistinguishable functions and 50%-identical amino acid sequences to determine whether the differences between the two proteins at the amino-acid level are functionally significant [23]. Their high degree of similarity assures very reliable sequence alignment, which establishes pairwise correspondence (based on aligned positions) for nearly all amino acid residues.[10] If aligned but non-matching residues are part-for-part equivalents, then we should be able to substitute freely among these equivalent pairs without impairment. Yet when protein sequences were even partially scrambled in this way, such that the hybrids were about 90% identical to one of the parents, none of them had detectable function. Considering the sensitivity of the functional test, this implies the hybrids had less than 0.1% of normal activity [23]. So part-for-part equiva-

[10] Of 257 positions in the alignment, two have gaps caused by insertion or deletion of a single amino acid [23].

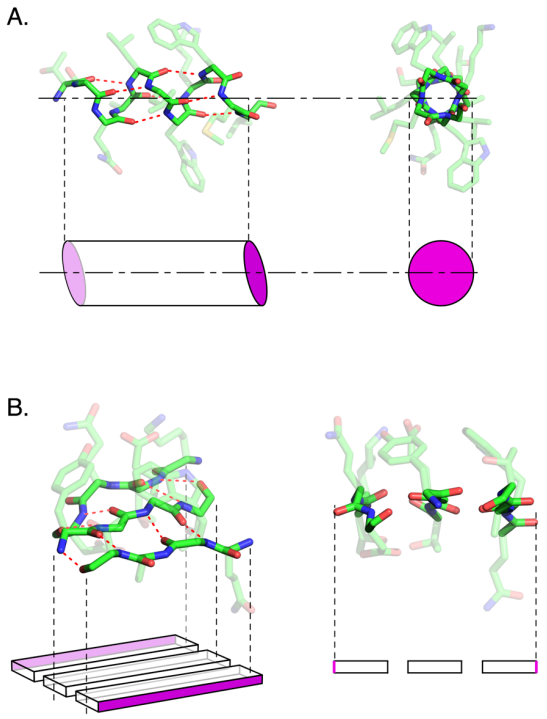

Figure 7. The restricted possibilities for 'generic' binding to alpha helices and beta sheets. Stick representations show backbones as solid and side chains as semi-transparent, with green, blue, and red representing carbon, nitrogen, and oxygen. Dashed red lines show the regular patterns of hydrogen bonding that stabilize secondary structure. Schematic diagrams beneath the molecular representations show (in simplified outline) where backbone atoms are accessible (purple), thereby allowing the secondary structure to be extended by continuation of the regular hydrogen bonding. A) In the standard alpha helix, shown from side and end, side chains protrude radially, making the exterior surface highly sequence dependent except at the exposed ends (purple), where the helix may be extended. B) In the standard beta sheet structures (parallel, or anti-parallel as shown) side chains protrude from both faces perpendicular to the plane of the sheet, making the exterior surface highly sequence dependent except at the exposed edges (purple), where the sheet may be extended. Right view shows sheet from strand ends.
doi:10.5048/BIO-C.2010.1.f7

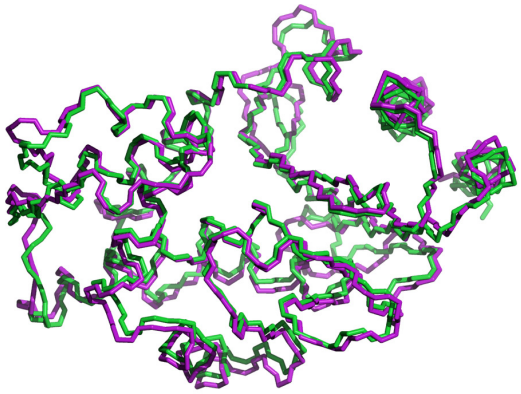

Figure 8. Aligned backbone structures of the TEM-1 and PSE-4 beta lactamases. The modules constructed by Meyer and co-workers [50] derive from these two beta lactamase variants (TEM-1 shown green, from PDB entry 1BTL; PSE-4 shown purple, form PDB entry 1G68) and from the SED-1 variant, for which no structure is available. Structures were aligned by the combinatorial extension method [51] (see http://cl.sdsc.edu/).
doi:10.5048/BIO-C.2010.1.f8

lence is not borne out at the level of amino acid side chains.

In view of the dominant role of side chains in forming the binding interfaces for higher levels of structure, it is hard to see how those levels can fare any better. Recognizing the non-generic nature of side chain interactions, Voigt and co-workers developed an algorithm that identifies portions of a protein structure that are most nearly self-contained in the sense of having the fewest side-chain contacts with the rest of the fold [49]. Using that algorithm, Meyer and co-workers constructed and tested 553 chimeric proteins that borrow carefully chosen blocks of sequence (putative modules) from any of three natural beta lactamases [50]. They found numerous functional chimeras within this set, which clearly supports their assumption that modules have to have few side chain contacts with exterior structure if they are to be transportable.

At the same time, though, their results underscore the limitations of structural modularity. Most plainly, the kind of modularity they demonstrated is not the robust kind that would be needed to explain new protein folds. The relatively high sequence similarity (34–42% identity [50]) and very high structural similarity of the parent proteins (Figure 8) favors successful shuffling of modules by conserving much of the overall structural context. Such conservative transfer of modules does not establish the robust transportability that would be needed to make new folds. Rather, in view of the favorable circumstances, it is striking how low the success rate was. After careful identification of splice sites that optimize modularity, four out of five tested chimeras were found to be completely non-functional, with only one in nine being comparable in activity to the parent enzymes [50]. In other words, module-like transportability is unreliable even under extraordinarily favorable circumstances. Although the limited transportability that did occur was enough for the authors to achieve their intended aim of generating sequence diversity [50], their results underscore the implausibility of the robust structural modules of interest here.

Graziano and co-workers have tested robust modularity directly by using amino acid sequences from natural alpha helices, beta strands, and loops (which connect helices and/or strands) to construct a large library of gene segments that provide these basic structural elements in their natural genetic contexts [52]. For those elements to work as robust modules, their structures would have to be effectively context-independent, allowing them to be combined in any number of ways to form new folds. A vast number of combinations was made by random ligation of the gene segments, but a search through 10^8 variants for properties that may be indicative of folded structure ultimately failed to identify any folded proteins. After a definitive demonstration that the most promising candidates were not properly folded, the authors concluded that "the selected clones should therefore not be viewed as 'native-like' proteins but rather 'molten-globule-like'" [52], by which they mean that secondary structure is present only transiently, flickering in and out of existence along a compact but mobile chain. This contrasts with native-like structure, where secondary structure is locked-in to form a well defined and stable tertiary fold. Their finding accords well with what we should expect in view of the above considerations. Indeed, it would be very puzzling if secondary structure *were* modular.

In fact, although whole structural domains may be self-contained in the sense of carrying complete information for their own folding, even *they* may fail to meet the second criterion for structural modularity given above, simply because they do not have generic exteriors. I describe here an experimental demonstration of this that was performed years ago but not previously reported. Again it uses beta lactamases, which are an attractive model system because of the abundance of published structures and the ease of measuring their activity in vivo. This test used the two natural beta lactamases shown in Figure 9, which have highly similar backbone structures despite the fact that their sequences match at only 26% of aligned positions. Both structures consist of two domains, the larger of which was referred to previously (Figure 5B). Sections of the two genes were recombined to encode a chimeric protein that combines the domains colored green and red in Figure 9. The overall structural and functional similarity of the parent enzymes suggests that this kind of domain recombination should work. But the non-generic nature of the interface between the two domains in combination with the substantial sequence dissimilarity indicates otherwise—a point confirmed by the lack of detectable function for the chimeric construct.

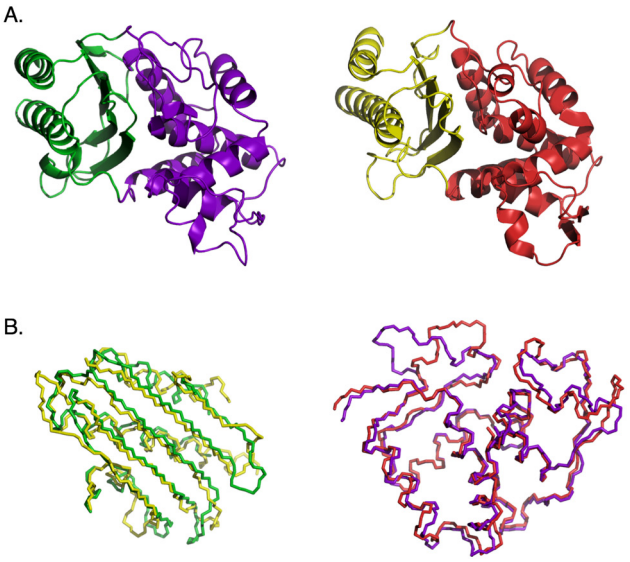

Figure 9. Structural comparison of beta lactamases used in construction of a domain chimera. A) The TEM-1 enzyme (left; PDB entry 1BTL) and the PER-1 enzyme (right; PDB entry 1E25) shown in same orientation with structural domains indicated by color. B) Backbone alignments for both domains, colored as in A. Structures were aligned by the combinatorial extension method [51]. The amino acid sequence of the chimeric construct, aligned with the parent sequences, is available as a supplementary file (**doi:**10.5048/BIO-C.2010.1.s1)
doi:10.5048/BIO-C.2010.1.f9

Pervasiveness of the problem

With no discernable shortcut to new protein folds, we conclude that the sampling problem really is a problem for evolutionary accounts of their origins. The final thing to consider is how pervasive this problem is. How often in the history of life would new phenotypes have required new protein folds? Or, narrowing that question, how much structural novelty do metabolic innovations appear to have required in the history of bacteria? Continuing to use protein domains as the basis of analysis, we find that domains tend to be about half the size of complete protein chains (compare Figure 10 to Figure 1), implying that two domains per protein chain is roughly typical. This of course means that the space of sequence possibilities for an average domain, while vast, is nowhere near as vast as the space for an average chain. But as discussed above, the relevant sequence space for evolutionary searches is determined by the combined length of *all* the new domains needed to produce a new beneficial phenotype.

As a rough way of gauging how many new domains are typically required for new adaptive phenotypes, the SUPERFAMILY database [54] can be used to estimate the number of different

protein domains employed in individual bacterial species, and the EcoCyc database [10] can be used to estimate the number of metabolic processes served by these domains. Based on analysis of the genomes of 447 bacterial species[11], the projected number of different domain structures per species averages 991[(12)]. Comparing this to the number of pathways by which metabolic processes are carried out, which is around 263 for *E. coli*,[13] provides a rough figure of three or four new domain folds being needed, on average, for every new metabolic pathway[14]. In order to accomplish this successfully, an evolutionary search would need to be capable of locating sequences that amount to anything from one in 10^{159} to one in 10^{308} possibilities[15], something the neo-Darwinian model falls short of by a very wide margin.

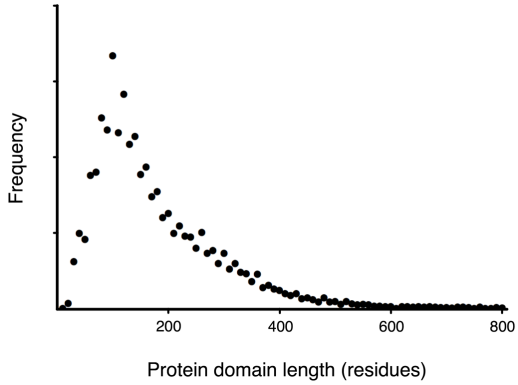

Figure 10. Length distribution for 9,535 SCOP-defined protein domains. The distribution is based on a non-redundant set (less than 40% pairwise sequence identity) obtained from ASTRAL SCOP version 1.73 [53]. The mean and median lengths are 178 residues and 145 residues, respectively. Lengths were binned in 10-residue increments, the most occupied bin containing 668 protein domains.
doi:10.5048/BIO-C.2010.1.f10

CONCLUSIONS

What was raised decades ago as an apparent limitation to the evolution of new proteins has here been dubbed the *sampling problem*—the impossibility of any evolutionary process sampling anything but a miniscule fraction of the possible protein sequences. At that time, several missing pieces of information made it difficult to conclude with certainty whether this limitation presented a serious challenge to neo-Darwinian accounts of the origin of new proteins. I have argued that the wealth of information that has accumulated since then has greatly strengthened the case that the sampling problem is real and that it does present a serious challenge.

We have used a picture of gems hidden in a vast desert at various points in our discussion in order to illustrate the challenge. Now that we have estimated the relevant fractions it may be helpful to return to this picture. Imagine that the search for gems is conducted by specifying sample points as mathematically exact geographic coordinate pairs (longitude and latitude). Sampling then consists of determining whether a gemstone rests at any of these specified points. A target the size of a grain of sand amounts to about one part in 10^{20} of a search space the size of the Sahara, which is above the feasibility threshold of one part in 5×10^{23}. So under favorable circumstances a Darwinian search would be capable of locating a sand-grain-sized gemstone in a Sahara-sized search space. As mentioned above, the ability to accomplish a search on this scale is clearly of some practical significance.

But as a generator of new protein folds, it turns out to be decidedly insignificant. Extending our desert picture, imagine that the top surface of every grain of sand in the Sahara has a miniature desert of its own resting upon it—one in which the entire Sahara is replicated in minute detail. We may call the sub-microscopic sand in these miniature deserts *level-1* sand, referring to the fact that it is one level removed from the real world (where we find *level-0* sand). This terminology can be applied to arbitrarily small targets by invoking a succession of levels (along the lines of De Morgan's memorable recursion of fleas[16]). In terms of this picture, the sampling problem stems from the fact that the targets for locating new protein folds appear to be much smaller than a grain of level-0 sand. For example, the target that must be hit in order to discover one new functional domain fold of typical size is estimated to cover not more than one ten-trillionth of the surface of a single grain of level-1 sand.[17] Under favorable circumstances a Darwinian search will eventually sample the grain of level-0 sand on which the right grain of level-1 sand rests, but even then the odds of sampling that level-1 grain are negligible, to say nothing of the target region on that grain.[18] And the situation rapidly deteriorates when we consider more relevant targets, like beneficial new phenotypes that employ (typically) several new protein structures. In the end, it seems that a search mechanism unable to locate a small patch on a grain of level-14 sand is not apt to provide the explanation of fold origins that we seek.[19]

Clearly, if this conclusion is correct it calls for a serious rethink of how we explain protein origins, and that means a rethink of biological origins as a whole. Drawing on some of the points developed here, I presented an earlier version of this case several years ago to two prominent experts in the field. Bothered by my conclusion, both felt that it must be in error. When the three of us met for a discussion, they had their own hunches about where my reasoning might have gone wrong. Interestingly, though, after perhaps two hours of heated discussion neither agreed with the other's hunch, and we ended up at a polite but dissatisfying impasse. I left with the distinct impression that my conclusion was being rejected not because it was unfounded but because it was unwelcome.

Many others may have had that impression after drawing similar conclusions in the decades since the birth of molecular biology. Whichever way the matter is ultimately resolved, everyone with a genuine interest in science should agree that there *is* a scientific case against the neo-Darwinian explanation of biological origins, the arguments put forward here representing only a part of that case. Like all scientific cases, this one will be judged by the evidence, and the diversity of opinion as to the outcome is, on the whole, a good thing for science. For those who continue to think that protein origins can be explained within a broadly Darwinian framework, it should now be clear what lines of evidence stand in the way of that for the rest of us.

[11] From SUPERFAMILY 1.69 release [54].

[12] Calculated by dividing the number of superfamilies detected in each species by the fraction of that species' genome assigned to any superfamily, then taking the mean of this ratio.[13] This is the number of metabolic pathways in the 12.5 release of EcoCyc, excluding "superpathways" to avoid redundancy.

[13] This is the number of metabolic pathways in the 12.5 release of EcoCyc, excluding "superpathways" to avoid redundancy.

[14] This is broadly consistent with the limited functional diversity of protein domain folds reflected in the SCOP classification [30], where superfamilies (structurally defined) contain an average of only 1.7 families (functionally defined).

[15] The higher prevalence is based on the chorismate mutase data [24] with $\ell = 153$ for each of three domains; the lower is based on the beta-lactamase data [25] with $\ell = 153$ for each of four domains.

[16] "Great fleas have little fleas upon their backs to bite 'em, And little fleas have lesser fleas, and so *ad infinitum*. ..." (Augustus De Morgan).

[17] Based on the chorismate mutase data [24] and a domain of average size ($\ell = 153$).

[18] The target might actually be fragmented into dots that appear on many different grains. Nonetheless it is the total target size in comparison to the total search space that determines the difficulty of the search.

[19] Based on the one in 10^{308} figure, from the beta-lactamase data [25] with $\ell = 153$ for each of four domains.

1. Eden M (1967) Inadequacies of neo-Darwinian evolution as a scientific theory. In: Moorhead PS, Kaplan MM, eds. Mathematical Challanges to the Neo-Darwinian Interpretation of Evolution. Philadelphia: Wistar Institute Press. pp 109-111.
2. Spetner LM (1968) Information transmission in evolution. IEEE Transactions on Information Theory. IT-14: 3-6. **doi:**10.1109/TIT.1968.1054070
3. Salisbury FB (1969) Natural selection and the complexity of the gene. Nature 224: 342-343. **doi:**10.1038/224342a0
4. Spetner LM (1970) Natural selection versus gene uniqueness. Nature 226: 948-949. **doi:**10.1038/226948a0
5. Berman HM, Goodsell DS, Bourne PE (2002) Protein structures: From famine to feast. Am Sci 90: 350-359. **doi:**10.1511/2002.4.350
6. http://www.rcsb.org/pdb/home/home.do
7. Calvin M, Benson AA (1948) The path of carbon in photosynthesis, Science 107: 476-480. **doi:**10.1126/science.107.2784.476
8. Krebs HA, Johnson WA (1937) The role of citric acid in intermediate metabolism in animal tissues. Enzymologia 4: 148-156.
9. Meyerhof O (1951) Mechanisms of glycolysis and fermentation. Can J Med Sci 29: 63-77.
10. http://ecocyc.org
11. Dembski WA (1998) The Design Inference. Cambridge: Cambridge University Press. p 209.
12. Gong S, Park C, Choi H, Ko J, Jang I, et al. (2005) A protein domain interaction interface database: InterPare. BMC Bioinformatics 6: 207. **doi:**10.1186/1471-2105-6-207
13. Chelikani P, Carpena X, Fita I, Loewen PC (2003) An electrical potential in the access channel of catalases enhances catalysis. J Biol Chem 278: 31290-31296. **doi:**10.1074/jbc.M304076200
14. Thoden JB, Huang X, Kim J, Raushel FM, Holden HM (2004) Long-range allosteric transitions in carbamoyl phosphate synthetase. Protein Sci 13: 2398-2405. **doi:**10.1110/ps.04822704
15. Huang X, Holden HM, Raushel FM (2001) Channeling of substrates and intermediates in enzyme-catalyzed reactions. Annu Rev Biochem 70: 149-180. **doi:**10.1146/annurev.biochem.70.1.149
16. Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. Nat Struct Biol 2: 856-863. **doi:**10.1038/nsb1095-856
17. Axe DD, Foster NW, Fersht AR (1996) Active barnase variants with completely random hydrophobic cores. P Natl Acad Sci USA 93: 5590–5594. **doi:**10.1073/pnas.93.11.5590
18. Keefe AD, Szostak JW (2001) Functional proteins from a random sequence library. Nature 410: 715-718. **doi:**10.1038/35070613
19. Yamauchi A, Nakashima T, Tokuriki N, Hosokawa M, Nogamai H, Arioka S, Urabe I, Yomo T (2002) Evolvability of random polypeptides through functional selection within a small library. Protein Eng 15: 619-626. **doi:**10.1093/protein/15.7.619
20. Hayashi Y, Sakata H, Makino Y, Uraba I, Yomo T (2003) Can an arbitrary sequence evolve towards acquiring a biological function? J Mol Evol 56: 162-168. **doi:**10.1007/s00239-002-2389-y
21. Yockey HP (1977) On the information content of cytochrome c. J Theor Biol 67: 345–376. **doi:**10.1016/0022-5193(77)90043-1
22. Reidhaar-Olson JF, Sauer RT (1990) Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. Proteins 7: 306–316. **doi:**10.1002/prot.340070403
23. 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
24. Taylor SV, Walter KU, Kast P, Hilvert D (2001) Searching sequence space for protein catalysts. P Natl Acad Sci USA 98: 10596-10601. **doi:**10.1073/pnas.191159298
25. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. J Mol Biol 341: 1295-1315. **doi:**10.1016/j.jmb.2004.06.058
26. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401-1404. **doi:**10.1126/science.1089370
27. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667-1686.
28. Axe DD, Foster NW, Fersht AR (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. Biochemistry-US 37: 7157-7166. **doi:**10.1021/bi9804028
29. Palzkill T, Le QQ, Venkatachalem KV, LaRocco M, Ocera H (1994) Evolution of antibiotic resistance: Several different amino acid substitutions in an active site loop alter the substrate profile of beta-lactamase. Mol Microbiol 12: 217-229. **doi:**10.1111/j.1365-2958.1994.tb01011.x
30. http://scop.mrc-lmb.cam.ac.uk/scop/
31. 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
32. Siew N, Fischer D (2003) Twenty thousand ORFan microbial protein families for the biologist? Structure 11: 7-9. **doi:**10.1016/S0969-2126(02)00938-3
33. Bashton M, Chothia C (2007) The generation of new protein functions by the combination of domains. Structure 15: 85-99. **doi:**10.1016/j.str.2006.11.009
34. Webber C, Barton GJ (2001) Estimation of P-values for global alignments of protein sequences. Bioinformatics 17: 1158-1167. **doi:**10.1093/bioinformatics/17.12.1158
35. Chou KC (1994) Prediction of protein folding types from amino acid composition by correlation angles. Amino Acids 6: 231-246. **doi:**10.1007/BF00813744
36. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. P Natl Acad Sci USA **doi:**10.1073/pnas.0906408106
37. de Chateau M, Holst E, Björck L (1996) Protein PAB, an albumin-binding bacterial surface protein promoting growth and virulence. J Biol Chem 271: 26609-26615. **doi:**10.1074/jbc.271.43.26609
38. He Y, Chen Y, Alexander P, Bryan PN, Orban J (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. P Natl Acad Sci USA 105: 14,412-14417. **doi:**10.1073/pnas.0805857105
39. Constance JJ, Gloss LM, Petsko GA, Ringe D (2000) The role of residues outside the active site: structural basis for function of C191 mutants of *Escherichia coli* aspartate aminotransferase. Protein Eng 13: 105-112. **doi:**10.1093/protein/13.2.105
40. Tsuji T, Kobayashi K, Yanagawa H (1999) Permutation of modules or secondary structure units creates proteins with basal enzymatic properties. FEBS Lett 453: 145-150. **doi:**10.1016/S0014-5793(99)00711-5
41. Bogarad LD, Deem MW (1999) A hierarchical approach to protein molecular evolution. P Natl Acad Sci 96: 2591-2595. **doi:**10.1073/pnas.96.6.2591
42. Matsuura T, Ernst A, Plückthun A (2002) Construction and characterization of secondary structure modules. Protein Sci 11: 2631-2643. **doi:**10.1110/ps.0215102
43. Wright PE, Dyson HJ, Lerner RA (1988) Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. Biochemistry-US 27: 7167-7175. **doi:**10.1021/bi00419a001
44. Watters AL, Deka P, Corrent C, Callender D, Varani G, Sosnik T, Baker D (2007) The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell 128: 613-624. **doi:**10.1016/j.cell.2006.12.042
45. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. P Natl Acad Sci USA 37: 205-211. **doi:**10.1073/pnas.37.4.205
46. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. P Natl Acad Sci USA 37: 251-256. **doi:**10.1073/pnas.37.5.251
47. Alber T, Bell JA, Sun DP, Nicholson H, Wozniak JA, Cook S, Matthews BW (1988) Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability. Science 239: 631-635. **doi:**10.1126/science.3277275
48. Remaut H, Waksman G (2006) Protein–protein interaction through β-strand addition. Trends Biochem Sci 31: 436-444. **doi:**10.1016/j.tibs.2006.06.007
49. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Biol 9: 553-558. **doi:**10.1038/nsb805
50. Meyer MM, Hochrein L, Arnold FH (2006) Structure-guided SCHEMA recombination of distantly related β-lactamases. Protein Eng Des Sel 19: 563-570. **doi:**10.1093/protein/gzl045
51. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11: 739-747. **doi:**10.1093/protein/11.9.739
52. Graziano JJ, Liu W, Perera R, Geierstanger BH, Lesley SA, Schultz PG (2008) Selecting folded proteins from a library of secondary structural elements. J Am Chem Soc 130: 176-185. **doi:**10.1021/ja074405w
53. http://astral.berkeley.edu/
54. http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/