Protein Fold Evolution by Creative Destruction

Claudia Alvarez-Carreño[1,2]*, Rohan Gupta[2], Anton S. Petrov[1,2]*, and Loren Dean Williams[1,2]*

[1]NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA

[2]School of Chemistry and Biochemistry, Georgia Institute of Technology, 901 Atlantic Dr, Atlanta, GA 30332, USA

* Corresponding authors:
Claudia Alvarez-Carreño
Email: ccarreno6@gatech.edu
Anton S. Petrov
Email: anton.petrov@biology.gatech.edu
Loren Dean Williams
Email: loren.williams@chemistry.gatech.edu

Keywords: Domain, DNA replication, tandem repeat, translation

Abstract

Mechanisms by which new protein folds emerge and diverge pose central questions in biological sciences. Incremental mutation and step-wise adaptation explain relationships between topologically similar protein folds. However, the universe of folds is diverse and riotous, suggesting roles of more potent and creative forces. Sequence and structure similarity are observed between topologically distinct folds, indicating that proteins with distinct folds may share common ancestry.

We found evidence of common ancestry between three distinct β-barrel folds: OB, SH3 and cradle loop barrel (CLB). Here we present a mechanism of fold evolution that interconverts SH3, OB and CLB. This mechanism, which we call Creative Destruction, can be generalized to explain other examples of fold evolution including circular permutation. In Creative Destruction, an open reading frame duplicates or otherwise merges with another to produce a fused polypeptide. A merger forces two ancestral domains into a new sequence and spatial context. The fused polypeptide can explore folding landscapes that are inaccessible to either of the independent ancestral domains. However, the folding landscapes of the fused polypeptide are not fully independent of those of the ancestral domains. Creative Destruction is thus partially conservative in that a daughter fold would inherit some motifs from the ancestor folds. After a merger and refolding, adaptive processes such as mutation and loss of extraneous segments optimize the new daughter fold.

Introduction

Polypeptides collapse into well-defined globular structures, either spontaneously or with assistance from chaperones. At the base of the protein structure hierarchy, polypeptide chains form intramolecular hydrogen bonds within α-helices, β-sheets and loops (Bragg, et al. 1950; Pauling, et al. 1951). At the next level of the hierarchy, secondary structural elements can combine to form supersecondary structural elements such as β-α-β, β-hairpin or helix-turn-helix (Rao and Rossmann 1973; Taylor and Thornton 1984; Söding and Lupas 2003). At even higher levels of the structure hierarchy, secondary and supersecondary structural elements assemble with each other to from compact globular domains.

A domain is an independent and cooperative unit of intramolecular polypeptide assembly (Chothia and Finkelstein 1990; Chothia and Gerstein 1997; Rose, et al. 2006). Domains are characterized by folds (Figure 1), which are specific layouts and topologies of secondary structural elements (Chothia, et al. 1997). Nearly a thousand folds form the basis of hundreds of thousands of proteins (Chothia 1992), generating the vast universe of function (Chothia and Gerstein 1997; Efimov 1997; Levitt 2009).

As revealed by the evolution of the translation machinery (Kovacs, et al. 2017), the simplest and most ancient protein domains are built from a small set of supersecondary structures. Contemporary folds emerged from a funneled exploration; there is insufficient time and resources in the universe to find any given fold by random searching of sequence space (Koonin and Galperin 2003).

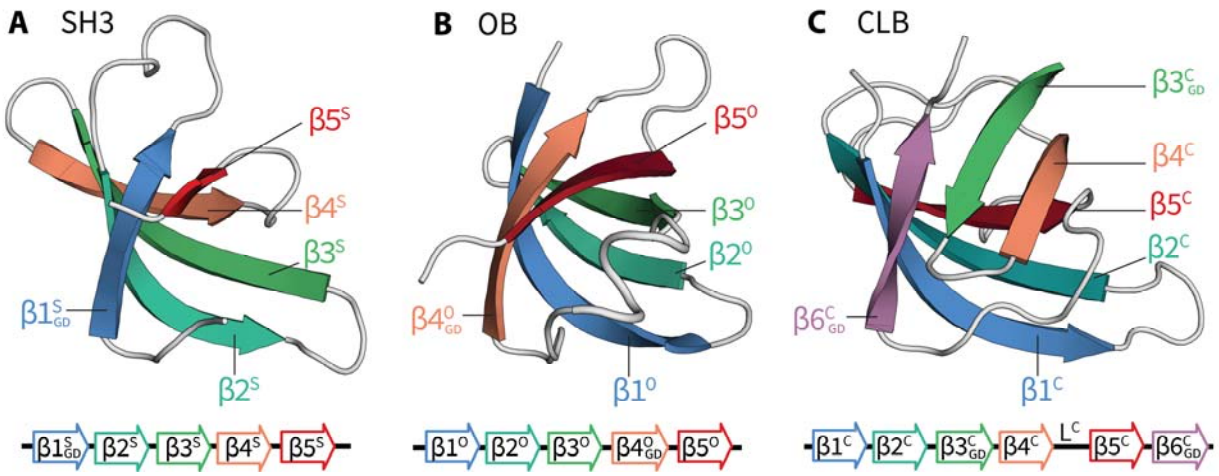

Figure 1. Representative structures of OB, SH3 and CLB folds. (A) Structure of an SH3 fold (PDB: 1NZ9, chain A). (B) Structure of an OB fold (PDB: 2OQK, chain A). (C) Structure of a CLB fold (PDB: 4B43, chain A). OB and SH3 are five-stranded β-barrels, CLB is a six-stranded β-barrel. Strands are colored as follows: β1 blue; β2 teal; β3 green; β4 orange; β5 red; β6 purple. GD: GD-box motif.

Mechanisms of fold origins and evolution pose central questions in biological sciences. How did ancient folds arise? What evolutionary mechanisms led to the broad diversity of protein folds in contemporary biological systems? Modest divergence from an ancestral fold can take place by incremental adaptive mutations that convert one type of secondary element to another (Grishin 2001). However small stepwise adaptive steps cannot explain large-scale folding rearrangements (Vogel and Morea 2006; Weiner and Bornberg-Bauer 2006).

Here we describe a general mechanism of fold evolution, called Creative Destruction, that appears to provide pathways for conversion between divergent classes of folds. Creative Destruction may account for the observed diversity of protein folds and affords experimental approaches to exploration of new fold space.

In Creative Destruction, an open reading frame duplicates or otherwise merges with another (Lauer and Gresham 2019) to produce a fused polypeptide. A merger forces two ancestral domains into a new sequence and spatial contexts, with the possibly of physical impingement. A merger thus has the potential to destabilize or partially destroy the folds of the ancestral gene products. At the same time the fused polypeptide might collapse to a new fold. The fused polypeptide can explore folding landscapes that are inaccessible to either of the independent ancestral domains. However, the folding landscapes of the fused polypeptide are not fully independent of those of the ancestral domains. Creative Destruction is thus partially

conservative in that a daughter fold would inherit some or all motifs from the ancestral folds and would also contain new elements. After gene merger, polypeptide expression, and exploration of new folding landscapes, adaptive processes such as mutation and loss of extraneous segments optimize the new daughter fold. In Creative Destruction, well-characterized and frequent genetic processes (Lauer and Gresham 2019) provide access to new folding landscapes.

In sum, Creative Destruction can create new folds that inherit some motifs from ancestral folds but whose overall topology is inaccessible by stepwise mutation of ancestral folds. Creative Destruction provides a general mechanism of fold diversification from a small subset of ancestral folds without requiring biologically implausible combinatorial shuffling of supersecondary structures between folds.

Here, to illustrate Creative Destruction, we describe its operation during circular permutation, in which two identical ancestral folds are destroyed. Circular permutation entails gene duplication (Jeltsch 1999; Grishin 2001), fold destruction and fold creation (in a new landscape), and adaptation. The secondary elements of the two ancestors are semiconserved in the daughter fold (half of them are conserved, Figure 2). In our model, circular permutation requires destruction of both ancestral folds, then folding in the context of a new landscape. The mechanism of circular permutation of the zinc-binding ribosomal protein uL33 (Kovacs, et al. 2018) illustrates the role of refolding and the conservation of secondary elements (Figure 2). Fifty percent of the secondary elements from the two ancestral domains are lost during circular permutation.

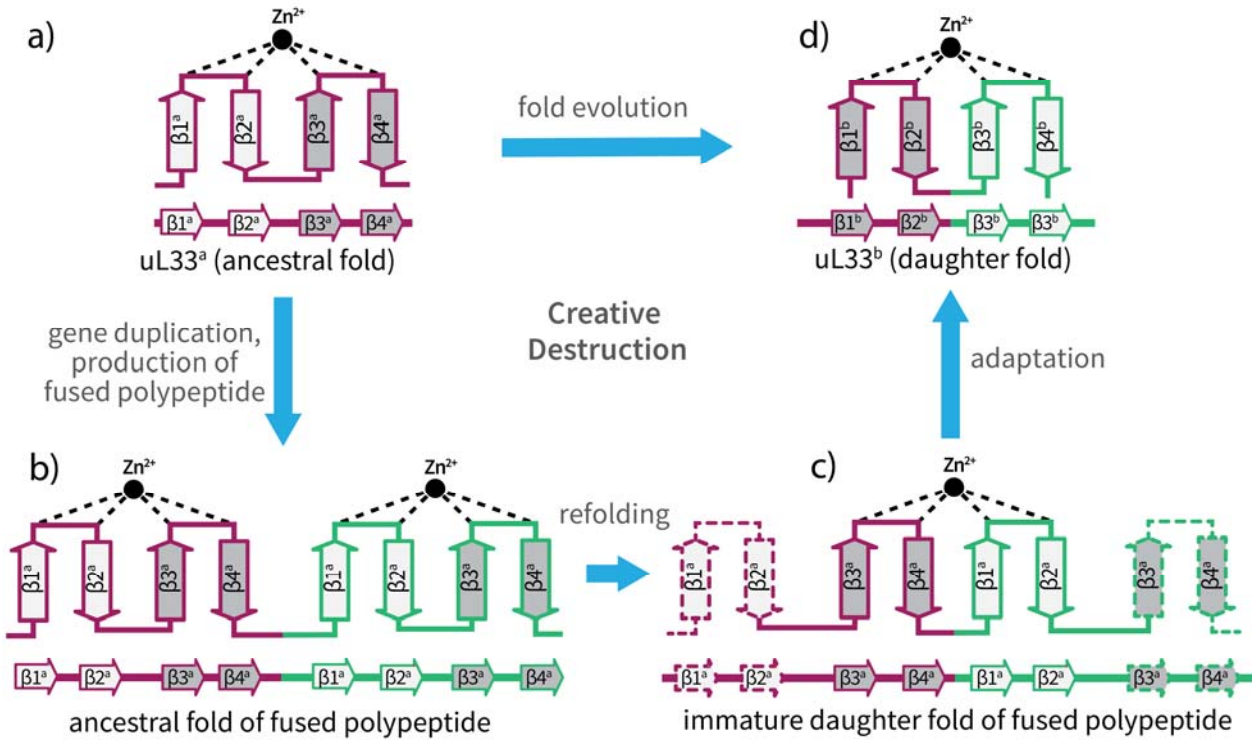

Figure 2. Circular permutation by Creative Destruction. An ancestral fold is destroyed, and a daughter fold is created. This figure shows, on a topological level, (a) the ancestral fold of archaeal uL33, (b) the ancestral folds of the fused polypeptide, (c) the immature daughter fold in the fused polypeptide, and (d) mature daughter fold of bacterial uL33. A duplication of $\beta1^a\beta2^a\beta3^a\beta4^a$ gives the fused polypeptide $\beta1^a\beta2^a$**$\beta3^a\beta4^a$** -- **$\beta1^a\beta2^a$**$\beta3^a\beta4^a$, (where -- is a linker). The circles represent zinc ions. The fused polypeptide refolds and resolves by adaption and the zinc ions shifts position or are released, yielding the bold secondary elements (above), which are renumbered $\beta1^b\beta2^b\beta3^b\beta4^b$ ($\beta1^a$ ~ $\beta3^b$, $\beta2^a$ ~ $\beta4^b$, $\beta3^a$ ~ $\beta1^b$, $\beta4^a$ ~ $\beta2^b$ where ~ indicates cross-fold sequence similarity). The dashed secondary elements are unstable in the immature daughter folds and are lost in the mature daughter fold. The ancestral folds of the fused polypeptide are included in the schematic to illustrate destruction of the ancestral folds and inherence of some ancestral secondary motifs. In reality the fused polypeptide would bypass the ancestral folds, and collapse directly to the immature daughter fold. $\beta3^a\beta4^a$ is shaded in dark gray to allow its path to be tracked during the transition from ancestral fold to daughter fold, where it is renumbered $\beta1^b\beta2^b$. The position of $\beta1^b\beta2^b$ is circularly permuted with respect to $\beta3^a\beta4^a$.

Here, we provide sequence and structural support for widespread occurrence of Creative Destruction during fold evolution. Our focus is on some of the oldest, simplest, and best-preserved proteins in biology. Vestiges of Creative Destruction are seen by comparisons of three ancient β-barrel folds: Oligonucleotide/Oligosaccharide-Binding, OB (Murzin 1993); Scr Homology 3 (SH3); and cradle loop barrel, CLB (Coles, et al. 2005). Proteins with OB, SH3 and CLB folds are found in central metabolic processes and throughout the translation system, including in ribosomal proteins, translation factors, and aminoacyl tRNA synthetases. We use CLB to refer to the Alanine Racemace C topology of the cradle loop barrel fold (Cheng, et al.

2014). Our results reveal Creative Destruction is a mechanism of circular permutation, and explain common ancestry of SH3, OB, and at least a subset of CLB fold proteins.

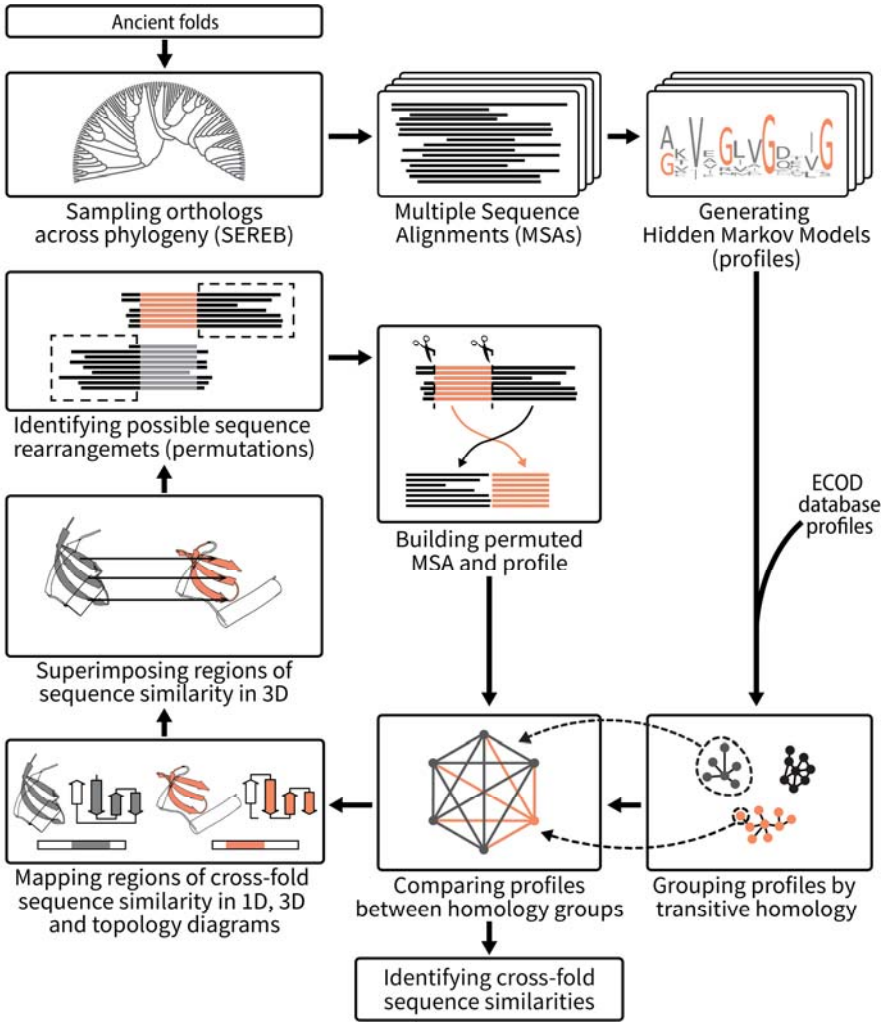

Figure 3. Workflow for identifying evolutionary relationships between folds. Orthologous protein sequences for a given fold within the SEREB database were aligned. Multiple sequence alignments (MSAs) were used to create profile Hidden Markov Models for each fold. Profile-profile comparisons between folds in the ECOD database revealed regions of sequence similarity. The method revealed transitive homologous relationships, in which two profiles are both similar to a third profiles (Roessler, et al. 2008). Transitive relationships between OB domains, SH3 domains and CLB domains analyzed here can be found in the Supplementary Materials. The method also revealed cross-fold similarities covering distinct and, in some cases, overlapping regions between distinct folds.

Results

Three-dimensional structures of biopolymers are generally more conserved than sequences (Illergård, et al. 2009; Bernier, et al. 2018). Exceptions to this pattern are called cross-fold sequence similarities, which are sequences that are conserved between proteins with different three-dimensional structures. Cross-fold sequence similarities indicate shared

evolutionary history between folds and are recognized as evidence of fold evolution (Grishin 2001; Söding and Lupas 2003; Farías-Rico, et al. 2014; Nepomnyachiy, et al. 2017). In some cases, different folds reveal multiple regions of sequence similarity that are permuted relative to each other.

We have detected homologous relationships within and between SH3, OB, and CLB folds. Our approach is based on detecting sequence similarities among protein profiles using the workflow in Figure 3 (see Methods). The HHalign score indicates the estimated probability of homology between a query and a template profile. Here, we consider HHalign scores higher than 60% to be indicative of shared common ancestry. In several instances, two sequences are seen to be homologous to each other by way of homology to a common third sequence (Park, et al. 1997). This phenomenon, called transitivity (Roessler, et al. 2008), has facilitated detection and characterization of signals of Creative Destruction. In other instances, sequence similarity is found between distinct folds (cross-fold sequence similarity). The patterns of cross-fold sequence similarities between distinct representatives of paired folds are consistent with, and provide strong support for, Creative Destruction as a mechanism of fold evolution.

We have also identified shared structural motifs between SH3, OB, and CLB, including, β-sheets, β-hairpins and GD-box motifs (Alva, et al. 2009; Alvarez-Carreño, et al. 2021). A GD-box is a motif that contains a β-strand connected to a loop by a β-turn, and portions of a second non-contiguous β-strand (Alva, et al. 2009). A GD-box is characterized in part by the amino acid sequence ΨxΨxxGρxΨxΨ, where G is glycine, Ψ is aliphatic, ρ is polar and x is anything.

Here, we support the Creative Destruction model by structure and sequence comparisons. To understand and explain Creative Destruction, we use a reduced representation of protein structure: 1) α refers to an α-helix, β refers to a β-strand, and L refers to a loop; 2) a number following an α or β indicates the relative position in the sequence, N to C; 3) secondary structural elements are notated by fold as α^O or β^O or L^O (indicates OB fold), α^S or β^S or L^S (indicates SH3), or α^C or β^C or L^C (indicates CLB); 4) L is only specified for loops that display cross-fold sequence similarity to an α-helix or a β-strand; 5) β_{GD} indicates a GD-box; 6) -- refers to a linker between two domains; and 6) ~ indicates cross-fold sequence similarity between

secondary elements. For example, $\beta1^{O} \sim \beta3^{S}$ indicates the first β-strand of an OB fold protein has sequence similarity with the third β-strand of an SH3 fold protein.

SH3 and OB Folds

OB and SH3 folds are both five-stranded β-barrels with two antiparallel β-sheets (Murzin 1993). Although the topology of SH3 and OB folds differs, β-strands and some of the connective elements overlay in three-dimensions (Agrawal and Kishan 2001; Alvarez-Carreño, et al. 2021). SH3 and OB folds share one region of cross-fold sequence similarity corresponding to $\beta2^{S}\beta3^{S}\beta4^{S} \sim \beta1^{O}\beta2^{O}\beta3^{O}$ (Figure 4A) and a second region of cross-fold sequence similarity corresponds to $\beta1_{GD}^{S}\ \beta2^{S} \sim \beta4_{GD}^{O}\ \beta5^{O}$ (Figure 4B).

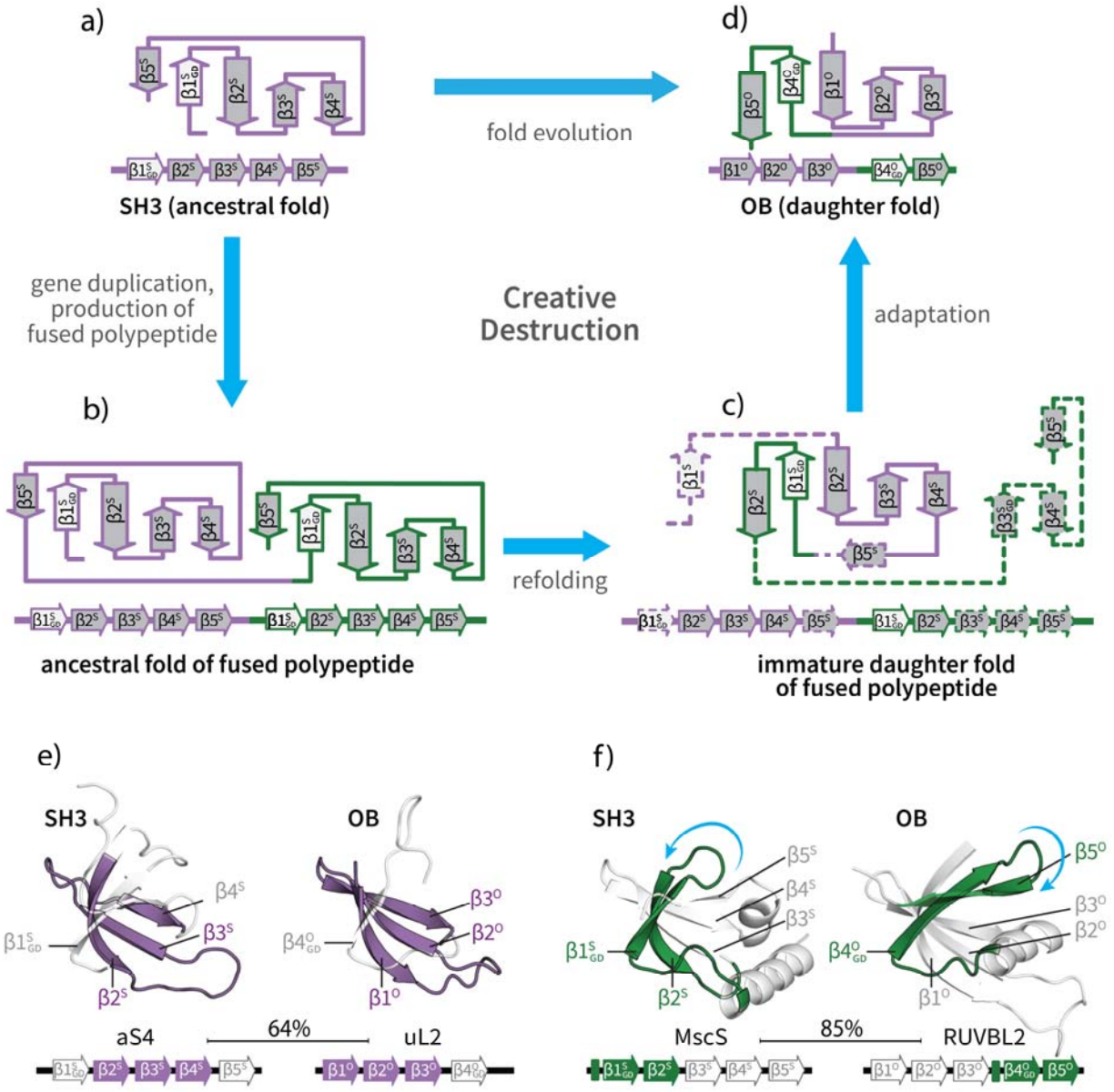

Figure 4. Conversion from SH3 Fold to OB Fold by Creative Destruction. This figure shows, on a topological level, (a) the ancestral fold of the native protein, (b) the ancestral folds of the fused polypeptide, (c) the immature daughter fold in the fused polypeptide, and (d) mature daughter fold. A duplication of SH3 gives the fused polypeptide SH3-SH3. The fused polypeptide refolds and resolves by adaption, yielding an OB fold. GD: GD-box motif. The ancestral folds of the fused polypeptide are included in the schematic to illustrate destruction of the ancestral folds and inherence of some ancestral secondary motifs. The fused polypeptide in reality would bypass the ancestral folds, and collapse directly to the immature daughter fold. $\beta2^S \sim \beta1^O$, $\beta2^O \sim \beta3^S$, $\beta4^S \sim \beta3^O$, and $\beta2^S \sim \beta5^O$ are shaded in dark gray allow tracking of β-strand positions during the transition from ancestral fold to daughter fold. (e) Cross-fold sequence similarity between aS4 and uL2 corresponds to antiparallel β-strands $\beta2^S\beta3^S\beta4^S$ and $\beta1^O\beta2^O\beta3^O$. (f) Cross-fold similarity between MscS and RUVBL2 corresponds to $\beta1_{GD}{}^S$ and $\beta4_{GD}{}^O$. Mapping of sequence similarity between SH3 and OB: A bar between two folds indicates cross-fold sequence similarity, as given by HHalign scores. Regions that yield no sequence similarity are shown in white. Secondary structural elements that share sequence similarity are indicated by the same color in both members of the pair. Curved arrows follow secondary structural elements that are display differences in conformation between folds.

The patterns of cross-fold sequence similarity, topology, motif conservation and 3D structure between SH3 and OB are consistent with Creative Destruction (Figure 4a-d). A duplication of $\beta 1_{GD}{}^{1}\beta 2^{S}\beta 3^{S}\beta 4^{S}\beta 5^{S}$ yields the fused polypeptide $\beta 1_{GD}{}^{S}\mathbf{\beta 2^{S}\beta 3^{S}\beta 4^{S}}\beta 5^{S}$ -- $\mathbf{\beta 1_{GD}{}^{S}\beta 2^{S}}\beta 3^{S}\beta 4^{S}\beta 5^{S}$. In our schematic, the fused polypeptide initially collapses to the ancestral folds, then converts to an immature daughter fold. The collapse to the ancestral folds is shown for illustrative purposes only; the fused polypeptide actually collapses directly to the immature daughter fold, which resolves by adaption. The final product contains bold secondary elements ($\mathbf{\beta 2^{S}\beta 3^{S}\beta 4^{S}\beta 1_{GD}{}^{S}\beta 2^{S}}$), which are renumbered (N to C) to give $\beta 1^{O}\beta 2^{O}\beta 3^{O}\beta 4_{GD}{}^{O}\beta 5^{O}$. Creative Destruction is supported by cross-fold sequence similarities $\beta 1^{O}$ ~ $\beta 2^{S}$, $\beta 2^{O}$ ~ $\beta 3^{S}$, $\beta 3^{O}$ ~ $\beta 4^{S}$, $\beta 4_{GD}{}^{O}$ ~ $\beta 1_{GD}{}^{S}$, $\beta 5^{O}$ ~ $\beta 2^{S}$. This pattern of cross-fold sequence similarity is consistent with formation of an OB fold by duplication of an SH3 gene, polypeptide expression, collapse to an immature daughter fold, and adaption, to yield a SH3 fold (Figure 4a-d).

The simpler topology of SH3 compared to OB suggests that SH3 is the ancestor of OB. However, the opposite polarity, where a duplicated, refolded and adapted OB results in the emergence of the SH3 fold is also consistent with the data. We previously described how gene fusion, refolding and adaptation provide a simple mechanism of conversion between OB and SH3 folds (Alvarez-Carreño, et al. 2021).

Ribosomal proteins uL2 and aS4

Cross-fold sequence similarity is seen between ribosomal protein uL2 (OB) and ribosomal protein aS4 (SH3) (HHalign score 64%). These two proteins share one region of sequence and structure similarity of the antiparallel β-sheet ($\beta 2^{S}\beta 3^{S}\beta 4^{S}$ ~ $\beta 1^{O}\beta 2^{O}\beta 3^{O}$, Figure 4e and Supplementary File 1, Table S4). Sequence similarity between uL2 (OB) and ribosomal protein aS4 (SH3) is not apparent in the antiparallel β-sheet that contains the N- and C-terminal strands.

RUVBL2 and MscS

Cross-fold sequence similarity is seen in RuvB-like AAA ATPase 2 (RUVBL2, OB) and the mechanosensitive channel of small conductance (MscS, SH3) (Figure 4f and Supplementary File 1, Table S4). The region of sequence similarity is $\beta1_{GD}^{S}\beta2^{S}$ ~ $\beta4_{GD}^{O}\beta5^{O}$ (HHalign score 85%).

As predicted by Creative Destruction conversion of SH3 to OB requires refolding. $\beta1_{GD}^{S}\beta2^{S}$ and $\beta4_{GD}^{O}\beta5^{O}$ are contained within different antiparallel β-sheets. $\beta5^{O}$ is on the edge of $\beta^{O}\beta^{O}\beta^{O}$ while $\beta2^{S}$ (similar in sequence to $\beta5^{O}$) is on the interior of $\beta^{S}\beta^{S}\beta^{S}$. The relative position of $\beta5^{O}$ is the same as $\beta5^{S}$ within the linear sequence, but not within the three-dimensional structure.

SH3/OB summary

Homology relationships between OB folds and SH3 folds can be generalized. We have shown that uL2 and aS4 ($\beta2^{S}\beta3^{S}\beta4^{S}$ ~ $\beta1^{O}\beta2^{O}\beta3^{O}$) as well as RUVBL2 and MscS ($\beta1_{GD}^{S}\beta2^{S}$ and $\beta4_{GD}^{O}\beta5^{O}$) are related via cross-fold sequence similarities. Additionally, we demonstrate transitive homology relationships: 1) between the OB fold of uL2 and the OB fold of RUVBL2 (Supplementary File 1, Table S1); and 2) between the SH3 fold of aS4 and the SH3 fold of MscS (Supplementary File 1, Table S2).

Thus, accounting for both cross-fold sequence similarity and transitive homology data, we can describe the general relationship between SH3 and OB folds as follows: 1) $\beta1^{O}\beta2^{O}\beta3^{O}$ is equivalent to $\beta2^{S}\beta3^{S}\beta4^{S}$; 2) $\beta4_{GD}^{O}\beta5^{O}$ is equivalent to $\beta1_{GD}^{S}\beta2^{S}$; 3) the GD-box motifs in $\beta4_{GD}^{O}$ and in $\beta1_{GD}^{S}$ are homologous; 4) a duplicated SH3 $\mathbf{\beta1_{GD}^{S}\beta2^{S}\beta3^{S}\beta4^{S}}\beta5^{S}\mathbf{-\beta1_{GD}^{S}\beta2^{S}}\beta3^{S}\beta4^{S}\beta5^{S}$, after adaptation results in the emergence of an OB fold. $\mathbf{\beta2^{S}\beta3^{S}\beta4^{S}\beta1_{GD}^{S}\beta2^{S}}$, renumbered N to C, gives $\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}^{O}\beta5^{O}$ (Figure 2d).

We previously observed that SH3 and OB folds share two conserved supersecondary structural motifs: $\beta1^{O}\beta2^{O}\beta3^{O}$ ~ $\beta2^{S}\beta3^{S}\beta4^{S}$ and $\beta4_{GD}^{O}$ ~ $\beta1_{GD}^{S}$ (Alvarez-Carreño, et al. 2021). These conserved motifs are circularly permuted in SH3 with respect to OB. Here, we identify cross-fold sequence similarity between $\beta5^{O}$ and $\beta2^{S}$, which are not structurally conserved. Structural variability of $\beta5^{O}$ as well as differences in structure between $\beta5^{O}$ and $\beta2^{S}$ can be attributed to refolding, as part of the Creative Destruction process (see Discussion).

OB and CLB Folds

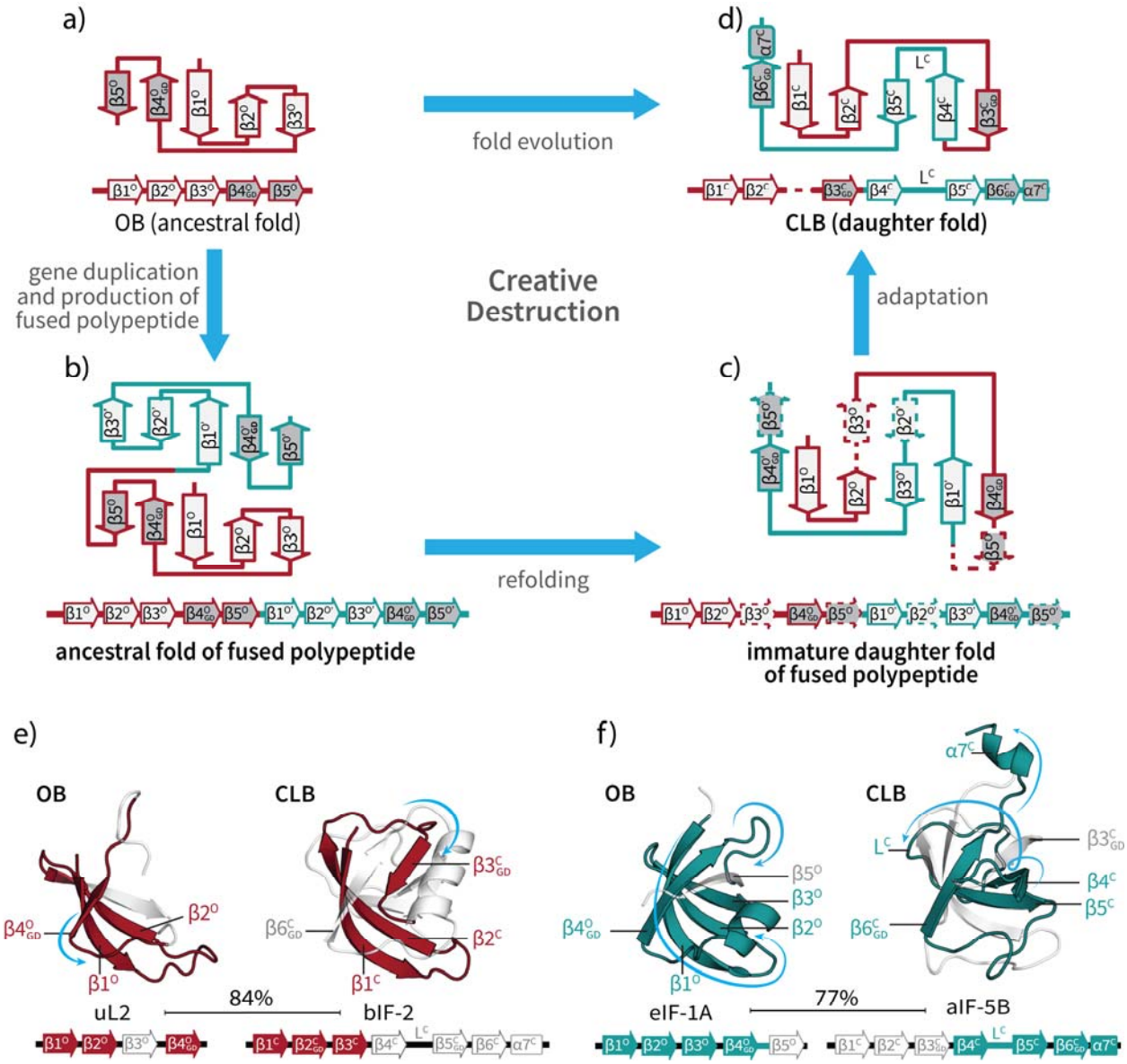

Figure 5. Conversion from OB Fold to CLB Fold by Creative Destruction. This figure shows, on a topological level, (a) the ancestral fold of the native protein, (b) the ancestral folds of the fused polypeptide, (c) the immature daughter fold in the fused polypeptide, and (d) mature daughter fold. A duplication of SH3 gives the fused polypeptide OB-OB. The fused polypeptide refolds and resolves by adaption, yielding a CLB fold. $\beta4^{O} \sim \beta3^{C}$, and $\beta4^{O} \sim \beta6^{C}$ are shaded dark gray to allow tracking of their positions during conversion from ancestral fold to daughter fold. (e) Cross-fold sequence similarity between uL2 and bIF2 corresponds to $\beta1^{o}\beta2^{o}\beta4_{GD}{}^{o}$ and $\beta1^{C}\beta2^{C}\beta3_{GD}{}^{C}$. (f) Cross-fold sequence similarity between eIF-1A and aIF-5B corresponds to $\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}{}^{O}L^{O}$ and $\beta4^{C}L^{C}\beta5^{C}\beta6_{GD}{}^{C}\alpha7^{C}$. Mapping of sequence similarity between OB and CLB: A bar between two folds indicates cross-fold sequence similarity, as given by HHalign scores. Regions that yield no sequence similarity are shown in white. Secondary structural elements that share sequence similarity are indicated by the same color in both members of the pair. Curved arrows follow secondary structural elements that are display differences in conformation between folds.

The CLB fold, with six β-strands, is larger and more complex than the OB fold; the number and linkages of β strands differ between these folds. However, commonalities between OB and CLB folds with Alanine Racemase C topology in ECOD suggest ancestry via Creative Destruction. Several distinct regions of some OB and CLB fold proteins show cross-fold sequence similarities and partially superimpose in three dimensions (Figure 5e and 5f). These folds share one region of cross-fold sequence similarity corresponding to $\beta1^{O}\beta2^{O}\beta4_{GD}^{O}$ and $\beta1^{C}\beta2^{C}\beta3_{GD}^{C}$, and a second region of cross-fold sequence similarity corresponding to $\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}^{O}L^{O}$ and $\beta4^{C}L^{C}\beta5^{C}\beta6_{GD}^{C}\alpha7^{C}$.

Shared features between OB and CLB folds are predicted by Creative Destruction. A duplicated and refolded OB fold, after adaptation, can give a CLB fold (Figure 5a-d). A duplication of $\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}^{O}\beta5^{O}$ gives the fused polypeptide $\mathbf{\beta1^{O}\beta2^{O}}\beta3^{O}\mathbf{\beta4_{GD}^{O}}\beta5^{O}$ -- $\mathbf{\beta1^{O'}\beta2^{O'}\beta3^{O'}\beta4_{GD}^{O'}}\beta5^{O'}$ (where the prime indicates copy 2). This conversion to CLB retains the bold secondary elements, which are renumbered $\beta1^{C}\beta2^{C}\beta3_{GD}^{C}\beta4^{C}L^{C}\beta5^{C}\beta6_{GD}^{O}\alpha7^{C}$ in the final product. The fused polypeptide refolds by extruding $\beta5^{O}$ and $\beta5^{O'}$. $\beta3^{O}$ is lost and $\beta2^{O'}$ refolds into a loop. $\beta4^{O'}$ forms the terminus of the β-sheet, next to $\beta1^{O}$. This model predicts cross-fold sequence similarities of $\beta1^{C} \sim \beta1^{O}$, $\beta2^{C} \sim \beta2^{O}$, $\beta3_{GD}^{C} \sim \beta4_{GD}^{O}$, $\beta4^{C} \sim \beta1^{O}$, $L^{C} \sim \beta2^{O}$, $\beta5^{C} \sim \beta3^{O}$, $\beta6_{GD}^{C} \sim \beta4_{GD}^{O}$, $\alpha7^{C} \sim L^{O}$.

Similarity between uL2 and bIF2

Cross-fold similarities are observed between the OB fold of uL2 and the CLB fold of bacterial Initiation factor 2 (bIF2) (HHalign score 84%). A region of similarity corresponds to $\beta1^{O}\beta2^{O}\beta4_{GD}^{O} \sim \beta1^{C}\beta2^{C}\beta3_{GD}^{C}$. $\beta4_{GD}^{O} \sim \beta3_{GD}^{C}$ have a homologous GD-box motif. In structure, a β-hairpin formed by $\beta1^{O}\beta2^{O}$ is similar to the β-hairpin formed by $\beta1^{S}\beta2^{S}$. Strand $\beta3^{S}$ has no equivalent in CLB.

Similarity between eIF-1A and aIF5B

Cross-fold similarity is detected between the entire OB fold of initiation factor eIF-1A and the CLB fold in aIF-5B (HHalign score 77%). Cross-fold sequence similarities between OB and CLB are: $\beta1^{O} \sim \beta4^{C}$, $\beta2^{O} \sim L^{C}$; $\beta3^{O} \sim \beta5^{C}$; $\beta4_{GD}^{O} \sim \beta6_{GD}^{C}$; $\beta3^{O}\beta4_{GD}^{O} \sim \beta5^{C}\beta6_{GD}^{C}$; and $L^{O} \sim \alpha7^{C}$. A GD-box sequence motif is located in the loops of $\beta4_{GD}^{O} \sim \beta6_{GD}^{C}$.

Ancestral relationship between OB and CLB folds

The CLB fold comprises a large group of folds that include six-stranded β-barrels with a variety of topologies (Ammelburg, et al. 2007; Alva, et al. 2008). Here we analyzed CLB folds of the Alanine Racemase C topology. The cross-fold sequence similarity of the OB fold can be mapped to two distinct regions of this CLB fold (Figures 5e, 5f). Transitive homology relationships within OB folds (Supplementary File 1, Table S1) and CLB folds with Alanine Racemase C topology (Supplementary File 1, Table S3) indicate that the three regions of cross-fold similarity between OB and CLB folds can be generalized as follows: 1) $\beta1^{C}\beta2^{C}\beta3_{GD}^{C}$ ~ $\beta1^{O}\beta2^{O}\beta4_{GD}^{O}$; 2) $\beta4^{C}L^{C}\beta5^{C}\beta6_{GD}^{C}\alpha7^{C}$ ~ $\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}^{O}L^{O}$; and 3) a duplicated four-stranded OB $\mathbf{\beta1^{O}}\beta2^{O}\mathbf{\beta3^{O}\beta4_{GD}^{O}}\beta5^{O}$--$\mathbf{\beta1^{O}\beta2^{O}\beta3^{O}\beta4_{GD}^{O}L^{O}}\beta5^{O}$ results in the CLB topology $\beta1^{C}\beta2^{C}\beta_{GD}3^{C}\beta4^{C}L^{C}\beta5^{C}\beta6_{GD}^{C}\alpha7^{C}$.

Discussion

In general, protein structure is more conserved than sequence (Illergård, et al. 2009). Sequence conservation between divergent folds is an indication of fold evolution (Grishin 2001). We identified sequence conservation (cross-fold sequence similarities) between three β-barrel folds: SH3, OB, and CLB. Here we present a mechanism of fold evolution that describes interconversions between SH3, OB and CLB.

Cross-fold sequence similarities between SH3 and OB folds suggest that the OB fold arose via duplication of an SH3 gene, followed by production and folding of a fusion protein, then adaptation, (Alvarez-Carreño, et al. 2021). Two SH3 genes fuse and collapse to a nascent OB fold that excludes terminal secondary elements. The mature OB fold was accomplished by loss of the terminal elements along with tuning mutations. OB fold proteins retain sequence and structure imprints from ancestral SH3 fold proteins. The polarity (SH3 -> OB versus OB -> SH3) is not fully resolved and does not affect the conclusion here. However, it seems most probable that SH3 is the ancestor of OB.

Cross-fold sequence similarities, identified here, suggest that OB gene duplication, protein refolding, and adaption produced the CLB fold. The mature CLB fold requires loss of

internal elements and tuning mutations. CLB fold proteins retain both sequence and structural imprints of OB fold proteins.

Creative Destruction

The combined data support a Creative Destruction model in which ancestral folds beget daughter folds. Each step of the mechanism is known to be accessible and relatively frequent. An initial gene fusion step causes insertion of one coding sequence in tandem with another (Lauer and Gresham 2019). It is assumed for simplicity that both ancestral genes express stable protein domains. A second step is expression of the fused polypeptide. The fusion alters the sequence and spatial context of the ancestral polypeptides and destabilizes the ancestral domains. At the same time alternative folding landscapes are opened. The fused polypeptide collapses to a daughter fold that differs from either ancestral fold. For clarity, the model here focuses on duplication of folding-competent domains. However, incorporation of non-coding sequences, incorporation of a genetic element within a coding sequence, or a truncation or partial deletion are expected to facilitate fold evolution.

Motif Inheritance

Fold evolution by Creative Destruction is a funnel, not a random search of fold space. Daughter folds are derived in partially conservative and easily accessible processes from ancestral folds. Daughter folds are contingent on ancestral folds and inherit sequence and structural elements from them.

The GD-box demonstrates surprising persistence between SH3, OB and CLB folds. The presence of these motifs appears to provide stabilization of the β-barrel folds by bringing together a β-turn and a non-contiguous β-strand (Alva, et al. 2009). SH3 and OB have one GD-box motif, and CLB has two (Figure 1). The Creative Destruction model explains the locations of these motifs in the sequence. Creative Destruction does not require complex and improbable shuffling processes or transfer of isolated folding-incompetent motifs between proteins, as suggested by some previous models of fold evolution (Alva, et al. 2015; Nepomnyachiy, et al. 2017; Kolodny, et al. 2021).

Plasticity

In a general scheme for Creative Destruction, a gene that encodes a protein with secondary elements ABCD can fuse with a gene that encodes a protein with secondary elements EFGH. The result is a fused polypeptide with ancestral secondary structure A**BC**D – E**FG**H. The fused polypeptide collapses and resolves by adaptation to daughter fold BCFG, which refolds to BCKG. The topology has been rewired, and F has changed conformation to K. Ancestral elements A, D, E, and H are lost. The most frequent conformational changes are expected to be conversion of α-helices to β-strands and vice versa (Grishin 2001). Some secondary elements of ancestors are retained while others are modified or lost. Conformational plasticity enables the integration of folding-competent sequences into new supersecondary structures.

Creative Destruction is seen here to explain patterns of conserved and non-conserved sequence and structure between OB, SH3 and CLB folds. Our data suggest conformational changes: $\beta 2^{O}$ converts to a loop in CLB (as exemplified by cross-fold sequence similarity between uL2 and bIF2, figure 5e), and a loop of the second copy of OB transforms to $\alpha 7^{C}$ (as demonstrated by cross-fold sequence similarity between eIF-1A and aIF5B, Figure 5f). We also note an absence of β5 in the daughter CLB fold compared to the ancestral OB fold. Similarly, β5 in the ancestral SH3 fold is absent from the daughter OB fold. It is possible that the ancestral folds were four-stranded. A second possibility is that the ancestors were five-stranded, but β5 was destroyed during the Creative Destruction process, which could have occurred either during the duplication and fusion event or at a later step during adaptation. The model provides a mechanism of conversion from one protein fold to another, and offers a basis for establishing evolutionary relationships among diverse folds.

Circular Permutation

Circular permutation has been documented previously in a variety of proteins including concanavalin A (Cunningham, et al. 1979), carbohydrate-binding proteins (Cunningham et al. 1979), saposins (Ponting and Russell 1995), DNA methyltransferases (Jeltsch 1999; Peisajovich, et al. 2006), and zinc ribbons (Ban, et al. 2000; Krishna, et al. 2003; Klein, et al. 2004; Kovacs, et al. 2018; Timsit, et al. 2021). Two circularly permuted domains differ only by two connections

between secondary elements, but are otherwise fully conserved. Circular permutation can arise by two distinct mechanisms. One mechanism occurs at the polypeptide level (Cunningham, et al. 1979) and is a direct post-translational re-wiring of the connections between secondary structural elements: 1) the N- and C-termini are joined; and 2) and a loop is cleaved, generating new N- and C-termini. Another mechanism occurs on the gene level (Jeltsch 1999), and as proposed here, involves Creative Destruction. Although the results appear indistinguishable, the two mechanisms are fundamentally different.

Circular permutation by Creative Destruction results from gene duplication, protein expression, exploration of new folding landscapes, and adaptation (Figure 2). The fused polypeptide product collapses to a new fold and adapts by loss of terminal elements. In circular permutation, the new folding landscape arrives at secondary elements inherited from the ancestral domains albeit with different linkages. Even though the ancestral domains and the daughter domain of circularly permuted proteins have common secondary structural elements, secondary elements of each ancestral domain are destroyed during the process.

We previously noted, for example, that archaeal and bacterial and versions of universal ribosomal protein uL33 are related by circular permutation (Figure 2) (Ban, et al. 2000; Klein, et al. 2004; Kovacs, et al. 2018). uL33 is a zinc ribbon, with two amphipathic β-hairpins (antiparallel ββ) linked by a zinc ion (Figure 2). The ordering of the ββ elements is switched in archaeal uL33 compared to bacterial uL33 ($\beta1^a\beta2^a$ ~ $\beta3^b\beta4^b$ and $\beta3^a\beta4^a$ ~ $\beta1^b\beta2^b$). Creative Destruction provides a simple mechanism of conversion of archaeal to bacterial uL33 (and vice versa). Duplication of $\beta1^a\beta2^a\beta3^a\beta4^a$ gives fused polypeptide $\beta1^a\beta2^a$**$\beta3^a\beta4^a$** -- **$\beta1^a\beta2^a$**$\beta3^a\beta4^a$. The fused protein collapses to a new fold, which omits ancestral elements $\beta1^a\beta2^a$ and $\beta3^a\beta4^a$, retaining the bold secondary elements to give daughter $\beta1^b\beta2^b\beta3^b\beta4^b$, where $\beta1^a$ ~ $\beta3^b$, $\beta2^a$ ~ $\beta4^b$, $\beta3^a$ ~ $\beta1^b$ and $\beta4^a$ ~ $\beta2^b$. The daughter fold adapts by loss of ancestral sequences for $\beta1^a\beta2^a$ and $\beta3^a\beta4^a$.

Fold evolution by Creative Destruction

In previous models of fold evolution, evolutionary transitions between folds have depended on persistent motifs (Lupas, et al. 2001; Söding and Lupas 2003; Alva, et al. 2015) or

themes (Nepomnyachiy, et al. 2017) that are smaller than domains and are ancestral to them. Proposed mechanisms of fold evolution include cooption of motifs and themes (Kolodny, et al. 2021; Longo, et al. 2022) and repetition and copy-pasting of motifs to generate new folds from small independent ancestral units (Zhu, et al. 2016). Creative Destruction by contrast acts on the level of domains and does not depend on assembly or folding of small elements outside the context of a domain. Creative Destruction does depend on fold plasticity, which is the facile exploration of new folding landscapes, and on a tendency of polypeptide to self-associate. Robust folding is seen experimentally in OB domain of uL2 and the CLB domain of bIF2 (To, et al. 2021). Creative Destruction provides a comprehensive and biologically plausible framework for the diversity of protein folds that does not require shuffling and interlacing of multiple fragments. Creative Destruction suggests that the universe of protein folds is better described as a network than as a tree (Nepomnyachiy, et al. 2014).

Methods

Representative SH3, OB and CLB domains of proteins from the translation system

Here we studied the evolutionary relationship between the most prevalent and simplest β-barrel folds observed within the translation machinery: OB, SH3 and CLB. The structural information for specific representative proteins (uL2, aS4, and aS28, bIF2, aIF-5B) that contain these folds is summarized in Table 1. MSAs of these representative proteins containing orthologous proteins from Sparse and Efficient Representation of Extant biology (Bernier 2018 MBE) were obtained from ProteoVision (Penev, et al. 2021). The MSAs of ribosomal proteins were trimmed to the domain boundaries as defined by the Evolutionary Classification of Domains, ECOD (Cheng, et al. 2014) and by Phase of Ribosomal evolution (Kovacs, et al. 2017). The MSAs for ribosomal protein domains (Supplementary Files 2-6) were transformed to HHMs (profile Hidden Markov Models) using the HHsuite version 3.3.0 (Steinegger, et al. 2019).

Finding transitive relationships and cross-fold sequence similarity

To understand the relationships of SH3, OB, and CLB within and across folds, we searched the ECOD database (Cheng, et al. 2014). HHM files were retrieved from ECOD v283 (http://prodata.swmed.edu/ecod/distributions/ecod.develop283.hhm_db). An initial sequence similarity search to identify sequence similarity was performed using HHsearch with our SH3, OB, and CLB HHMs as queries. To distinguish transitive homology relationships from cross-fold sequence similarities according to the hierarchical classification of ECOD the following criteria were applied: 1) HHalign scores greater than 60% within the same X-, H- and T-level groups in ECOD were considered transitive homologous relationships; 2) Protein domains yielding HHalign scores greater than 60% within different X-, H- or T-level groups in ECOD were considered cross-fold sequence similarities.

HHM files of domains displaying either transitive homologous relations (Supplementary File 1, Table S1-S3) or cross-fold similarities (Supplementary File 1, Table S4 and S5) were retrieved and were compared in an all-versus-all fashion using HHalign with default parameters (Steinegger, et al. 2019). HHalign scores of homology of pairwise comparisons were deposited in a 2x2 matrix (Supplementary File 1, Figure S1).

Structural analysis

Topology diagrams for these coordinate files were generated in ProteoVision using PDBsum (Laskowski, et al. 2018). Three-dimensional representations of specific folds in the selected PDB files (Table 1) were rendered in PyMol (Schrödinger_LLC 2021). Pairs of folds displaying cross-fold sequence similarity were superimposed using Click (Nguyen and Madhusudhan 2011) and manually adjusted using the pair fitting option in PyMol. Residues involved in cross-fold sequence similarities were highlighted by various colors in the topology diagrams and three-dimensional structure representations.

Table 1. Representative SH3, OB and CLB folds

Protein name	Fold Name	PDB code and chain	PDB fold range
uL2	OB	1VY4, chain BA	74-119
aS4	SH3	4V6U, chain AE	180-243
aS28	OB	4V6U, chain M	2-71
bIF-2	CLB	1ZO1, chain B	190-267
aIF-5B	CLB	4V8Z, chain EC	477-559

Data Availability

Sequence alignments associated with this manuscript have been deposited in the FigShare repository at https://figshare.com/s/aff5583321ad09dd7f3a.

Acknowledgements

The authors would like to thank Jessica Bowman and Dr. Eric Smith for insightful discussions. CA-C. was supported by the NASA Postdoctoral Program. This work was funded by the National Aeronautics and Space Administration grant 80NSSC18K1139 awarded to LDW and ASP.

References

Agrawal V, Kishan RK. 2001. Functional Evolution of Two Subtly Different (Similar) Folds. BMC Struct Biol 1:5.

Alva V, Dunin-Horkawicz S, Habeck M, Coles M, Lupas AN. 2009. The GD Box: A Widespread Noncontiguous Supersecondary Structural Element. Protein Sci 18:1961-1966.
Alva V, Koretke KK, Coles M, Lupas AN. 2008. Cradle-Loop Barrels and the Concept of Metafolds in Protein Classification by Natural Descent. Curr Opin Struct Biol 18:358-365.
Alva V, Söding J, Lupas AN. 2015. A Vocabulary of Ancient Peptides at the Origin of Folded Proteins. eLife 4:e09410.
Alvarez-Carreño C, Penev PI, Petrov AS, Williams LD. 2021. Fold Evolution before LUCA: Common Ancestry of Sh3 Domains and Ob Domains. Mol Biol Evol 38:5134-5143.
Ammelburg M, Hartmann MD, Djuranovic S, Alva V, Koretke KK, Martin J, Sauer G, Truffault V, Zeth K, Lupas AN, Coles M. 2007. A CTP-Dependent Archaeal Riboflavin Kinase Forms a Bridge in the Evolution of Cradle-Loop Barrels. Structure 15:1577-1590.
Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 a Resolution. Science 289:905-920.
Bernier CR, Petrov AS, Kovacs NA, Penev PI, Williams LD. 2018. Translation: The Universal Structural Core of Life. Mol Biol Evol 35:2065-2076.
Bragg WL, Kendrew JC, Perutz MF. 1950. Polypeptide Chain Configurations in Crystalline Proteins. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences 203:321-357.
Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. 2014. ECOD: An Evolutionary Classification of Protein Domains. PLoS Comput Biol 10:e1003926.
Chothia C. 1992. Proteins. One Thousand Families for the Molecular Biologist. Nature 357:543-544.
Chothia C, Finkelstein AV. 1990. The Classification and Origins of Protein Folding Patterns. Annu Rev Biochem 59:1007-1039.
Chothia C, Gerstein M. 1997. Protein Evolution. How Far Can Sequences Diverge? Nature 385:579, 581.
Chothia C, Hubbard T, Brenner S, Barns H, Murzin A. 1997. Protein Folds in the All-Beta and All-Alpha Classes. Annu Rev Biophys Biomol Struct 26:597-627.
Coles M, Djuranovic S, Söding J, Frickey T, Koretke K, Truffault V, Martin J, Lupas AN. 2005. Abrb-Like Transcription Factors Assume a Swapped Hairpin Fold That Is Evolutionarily Related to Double-Psi Beta Barrels. Structure 13:919-928.
Cunningham BA, Hemperly JJ, Hopp TP, Edelman GM. 1979. Favin Versus Concanavalin A: Circularly Permuted Amino Acid Sequences. Proc Natl Acad Sci USA 76:3218-3222.
Efimov AV. 1997. Structural Trees for Protein Superfamilies. Proteins 28:241-260.
Farías-Rico JA, Schmidt S, Höcker B. 2014. Evolutionary Relationship of Two Ancient Protein Superfolds. Nat Chem Biol 10:710-715.
Grishin NV. 2001. Fold Change in Evolution of Protein Structures. J Struct Biol 134:167-185.
Illergård K, Ardell DH, Elofsson A. 2009. Structure Is Three to Ten Times More Conserved Than Sequence—a Study of Structural Response in Protein Cores. Proteins: Struct Funct Bioinform 77:499-508.
Jeltsch A. 1999. Circular Permutations in the Molecular Evolution of DNA Methyltransferases. J Mol Evol 49:161-164.
Klein DJ, Moore PB, Steitz TA. 2004. The Roles of Ribosomal Proteins in the Structure Assembly, and Evolution of the Large Ribosomal Subunit. J Mol Biol 340:141-177.

Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. 2021. Bridging Themes: Short Protein Segments Found in Different Architectures. Mol Biol Evol.
Koonin EV, Galperin MY. 2003. Sequence-Evolution-Function: Computational Approaches in Comparative Genomics: Kluwer Academic, Boston.
Kovacs NA, Penev PI, Venapally A, Petrov AS, Williams LD. 2018. Circular Permutation Obscures Universality of a Ribosomal Protein. J Mol Evol 86:581-592.
Kovacs NA, Petrov AS, Lanier KA, Williams LD. 2017. Frozen in Time: The History of Proteins. Mol Biol Evol 34:1252-1260.
Krishna SS, Majumdar I, Grishin NV. 2003. Structural Classification of Zinc Fingers: Survey and Summary. Nucleic Acids Res 31:532-550.
Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. 2018. Pdbsum: Structural Summaries of PDB Entries. Protein Sci 27:129-134.
Lauer S, Gresham D. 2019. An Evolving View of Copy Number Variants. Curr Genet 65:1287-1295.
Levitt M. 2009. Nature of the Protein Universe. Proc Natl Acad Sci USA 106:11079-11084.
Longo LM, Kolodny R, McGlynn SE. 2022. Evidence for the Emergence of B-Trefoils by 'Peptide Budding' from an IgG-Like B-Sandwich. PLoS Comput Biol 18:e1009833.
Lupas AN, Ponting CP, Russell RB. 2001. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? J Struct Biol 134:191-203.
Murzin AG. 1993. Ob(Oligonucleotide/Oligosaccharide Binding)-Fold: Common Structural and Functional Solution for Non-Homologous Sequences. EMBO J 12:861-867.
Nepomnyachiy S, Ben-Tal N, Kolodny R. 2017. Complex Evolutionary Footprints Revealed in an Analysis of Reused Protein Segments of Diverse Lengths. Proc Natl Acad Sci USA 114:11703-11708.
Nepomnyachiy S, Ben-Tal N, Kolodny R. 2014. Global View of the Protein Universe. Proc Natl Acad Sci USA 111:11691-11696.
Nguyen MN, Madhusudhan MS. 2011. Biological Insights from Topology Independent Comparison of Protein 3D Structures. Nucleic Acids Res 39:e94.
Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate Sequences Increase the Detection of Homology between Sequences. J Mol Biol 273:349-354.
Pauling L, Corey RB, Branson HR. 1951. The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. Proc Natl Acad Sci USA 37:205-211.
Peisajovich SG, Rockah L, Tawfik DS. 2006. Evolution of New Protein Topologies through Multistep Gene Rearrangements. Nat Genet 38:168-174.
Penev PI, McCann HM, Meade CD, Alvarez-Carreño C, Maddala A, Bernier CR, Chivukula VL, Ahmad M, Gulen B, Sharma A, Williams LD, Petrov AS. 2021. ProteoVision: Web Server for Advanced Visualization of Ribosomal Proteins. Nucleic Acids Res 49:W578-w588.
Rao ST, Rossmann MG. 1973. Comparison of Super-Secondary Structures in Proteins. J Mol Biol 76:241-256.
Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, Montfort WR, Cordes MH. 2008. Transitive Homology-Guided Structural Studies Lead to Discovery of Cro Proteins with 40% Sequence Identity but Different Folds. Proc Natl Acad Sci USA 105:2343-2348.

Rose GD, Fleming PJ, Banavar JR, Maritan A. 2006. A Backbone-Based Theory of Protein Folding. Proc Natl Acad Sci USA 103:16623-16633.
Schrödinger_LLC, 2021. The Pymol Molecular Graphics System. Version 2.4.0
Söding J, Lupas AN. 2003. More Than the Sum of Their Parts: On the Evolution of Proteins from Peptides. Bioessays 25:837-846.
Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. Hh-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation. BMC Bioinformatics 20:473.
Taylor WR, Thornton JM. 1984. Recognition of Super-Secondary Structure in Proteins. J Mol Biol 173:487-512.
Timsit Y, Sergeant-Perthuis G, Bennequin D. 2021. Evolution of Ribosomal Protein Network Architectures. Sci Rep 11:625.
To P, Whitehead B, Tarbox HE, Fried SD. 2021. Nonrefoldability Is Pervasive across the E. Coli Proteome. J Am Chem Soc 143:11435-11448.
Vogel C, Morea V. 2006. Duplication, Divergence and Formation of Novel Protein Topologies. Bioessays 28:973-978.
Weiner J, 3rd, Bornberg-Bauer E. 2006. Evolution of Circular Permutations in Multidomain Proteins. Mol Biol Evol 23:734-743.
Zhu H, Sepulveda E, Hartmann MD, Kogenaru M, Ursinus A, Sulz E, Albrecht R, Coles M, Martin J, Lupas AN. 2016. Origin of a Folded Repeat Protein from an Intrinsically Disordered Ancestor. eLife 5.