Precise assembly of complex beta sheet topologies from de novo designed building blocks

Indigo Chris King, James Gleixner, Lindsey Doyle, Alexandre Kuzin, John F Hunt, Rong Xiao, Gaetano T Montelione, Barry L Stoddard, Frank DiMaio, David Baker

DOI: http://dx.doi.org/10.7554/eLife.11012

Cite as: eLife 2015;10.7554/eLife.11012

Received: 20 August 2015
Accepted: 8 December 2015
Published: 9 December 2015

Precise Assembly of Complex Beta Sheet Topologies from *de novo* Designed Building Blocks.

Indigo Chris King*[†], James Gleixner[†], Lindsey Doyle[§], Alexandre Kuzin[‡], John F. Hunt[‡], Rong Xiao[x], Gaetano T. Montelione[x], Barry L. Stoddard[§], Frank Dimaio[†], David Baker[†]

† Institute for Protein Design, University of Washington, Seattle, Washington, United States
‡ Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, New York, United States
[x] Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States
§ Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States

ABSTRACT

Design of complex alpha-beta protein topologies poses a challenge because of the large number of alternative packing arrangements. A similar challenge presumably limited the emergence of large and complex protein topologies in evolution. Here we demonstrate that protein topologies with six and seven-stranded beta sheets can be designed by insertion of one de novo designed beta sheet containing protein into another such that the two beta sheets are merged to form a single extended sheet, followed by amino acid sequence optimization at the newly formed strand-strand, strand-helix, and helix-helix interfaces. Crystal structures of two such designs closely match the computational design models. Searches for similar structures in the SCOP protein domain database yield only weak matches with different beta sheet connectivities. A similar beta sheet fusion mechanism may have contributed to the emergence of complex beta sheets during natural protein evolution.

INTRODUCTION

Modular domains constitute the primary structural and functional units of natural proteins. Multi-domain proteins likely evolved through simple linear concatenation of successive domains onto the polypeptide chain or through the insertion of one or more continuous sequences into the middle of another, now discontinuous domain [1-4]. By analogy, new proteins have been engineered from existing domains by simple linear concatenation or insertion of one domain into another [5-11]. How individual domains evolved, in contrast, is much less clear. Both experimental and computational analyses have suggested that new folds can evolve by insertion of one fold into another [3,12-14 15,16], but to our

knowledge there is no evidence that complex beta sheet topologies can be formed in this manner. On the protein design front, there has been progress in de novo design of idealized helical bundles [17] and alpha beta protein structures with up to 5 strands [18], and though new folds have been generated by tandem fusion of natural protein domains followed by introduction of additional stabilizing mutations[19,20], assembly of large and complex beta sheets poses a challenge for de novo protein design.

One possible route to the large and complex beta sheet topologies found in many native protein domains is recombination of two smaller beta sheet domains. Here we explore the viability of such a mechanism by inserting one de novo designed alpha beta protein into another such that the two beta sheets are combined into one. The backbone geometry at the junctions between the original domains is regularized, and the sequence at the newly formed interface is optimized to stabilize the single integrated domain structure. Crystal structures of two such proteins demonstrate that complex beta sheet structures can be designed with considerable accuracy using this approach, and provide a proof-of-concept for the hypothesis that complex beta topologies in natural proteins may have evolved from simpler beta sheet structures in a similar manner.

RESULTS

A first extended sheet protein was created by inserting a designed ferredoxin domain into a beta turn of the designed top7 protein to create a half-barrel structure, with the two sheets fused into a single seven strand sheet flanked by four helices (Figure 1A). The CD spectra show both alpha and beta structure (Figure 2—figure supplement 1). Two crystal structures (NESG target OR327) were solved by molecular replacement and refined to 2.49 Å (PDB entry 4KYZ) and 2.96 Å (PDB entry 4KY3) resolutions. Further analysis refers only to the higher resolution structure (4KYZ). The structure shows excellent agreement with the design model (Figure 2A), particularly in low B-factor regions, with C-alpha RMSD ranging from 1.76-1.85 Å among the four protomers in the crystal. The relative orientation of the strands packed against the helices is close to that in the design model, and core sidechains at the designed interfaces are in very similar conformations in the design model and crystal (Figure 2B,2C).

A second extended sheet protein was created by combining two designed ferredoxin domains via domain insertion to create a half-barrel structure with four alpha helices and six beta strands (Figure 1B). A beta turn segment between

two beta strands of the host ferredoxin was removed and the resulting cut-points in the host beta strands were linked to two beta strand cut-points in the insert, fusing the two strand pairs into a single, longer pair the center of a six-stranded beta sheet. CD spectra show that the protein contains both alpha and beta structure (Figure 3—figure supplement 1). Crystals were obtained which diffracted to 3.3Å resolution. Molecular replacement using the computational design models [21] yielded a solution for which the refinement statistics are shown in Supplementary File 1 (PDB entry 5CW9). Attempts to improve these statistics by rebuilding portions of the model proved unsuccessful, possibly due to a register shift or dynamic fluctuations in the structure (perhaps corresponding to slightly 'molten-globule'-like behavior) that are difficult to computationally model. However, unbiased low-resolution omit maps suggest that the overall topology is correct (Figure 3—figure supplement 2). In the model that displays the best refinement statistics, the protein backbone was similar to the design model with a C-alpha RMSD value of 2 Å (Figure 3A,3B). The fused beta sheet aligns with the design model, while the inter-domain helices shift slightly to accommodate the inter-domain interface. The sidechain packing between the newly juxtaposed beta strands succeeded in anchoring the secondary structure elements in their intended orientations, but the low resolution of the crystal structure prevents evaluation of the atomic-level accuracy of the design (Figure 3—figure supplement 2).

To compare the folds of these designed proteins to those in the SCOP v.1.75 domain database [22], the TMalign structure-structure comparison method was used to search a 70% sequence non-redundant set of SCOP domains [16] for structure alignments containing a minimum 75% overlap with the designed proteins. The most similar SCOP domains had weak TM-align scores (0.54 and 0.51) and the sheets in these matched structures have different connectivities than those of the designs, suggesting that the two designed proteins have novel folds (Figure 4). While there are no domains with globally similar folds, both designed proteins are similar to a number of SCOP domains over the ferrodoxin-like substructure(s) as is made evident by mapping the proteins to the domains network of Nepomnyachiy et al. [16] (Figure 4—figure supplement 1). The mutations introduced at the redesign stage of the domain insertion design protocol are compatible with the parent fold structures with minimal perturbation of the protein backbone (Figure 4—figure supplement 2) suggesting the designed folds would have the potential to evolve from insertion followed by neutral mutational drift of the parent structures.

DISCUSSION

We have shown that single designed protein domains can be combined into larger domains with complex beta sheet topologies. This mechanism provides a straightforward route to designing large and complex beta sheet structures capable of scaffolding the pockets and cavities essential for future design of protein functions. Our success in designing larger beta sheet domains by recombining smaller independently folded beta sheet proteins suggests a similar mechanism could have played a role in the evolution of naturally occurring complex beta sheet proteins.

MATERIALS AND METHODS

Our design strategy began with selection of three previously characterized de novo designed protein domains to serve as building blocks for recombination through domain insertion: ferredoxin, rossman 2x2, and top7 [18]. These three domains were chosen because they were the only Rosetta *de novo* designed protein domains with both alpha and beta secondary structure for which high resolution experimental structures had been obtained at the time of this work. Each chimeric domain consists of a parent host domain and a parent insert domain. In the insert domain, three residues from from the n-terminus were paired with three residue from the c-terminus to create nine residue pairs. Each residue pair was then aligned against all pairs of residues in the host domain to search for possible insertion points. Insertion points were accepted for residue pair alignment distances of 1 angstrom RMSD or less, replacing host domain segments of less than 5 residues. For every insertion point, a structure is generated by removing the residues between the insertion residues of the host domain and adding linkers between the aligned host and insert domain residues (Figure 1). Host and insert were connected by addition of 1-3 residues at the domain junctions using Rosetta Remodel [23], and 12 models in which this junction formed a continuous beta strand were identified. The sequences of these chimeras were optimized using Rosetta Design calculations around the junction regions and the new interface between the former domains. During the design simulation, all amino acid positions within 5 Å of the inter-domain junction interface were redesigned to minimize the predicted free energy of folding with the Rosetta all-atom energy function and a flexible backbone protein design protocol described previously [23]. Final designs were selected based on Rosetta energy, packing metrics, and similarity of the junction backbone geometry to local backbone geometry in the PDB. Twelve final domain insertion designs were chosen for expression in E. coli as 6xHis-tag fusions and purified on a Ni-NTA column. Purified proteins were evaluated for the presence of alpha/beta secondary structures via circular dichroism spectroscopy (CD), and three with levels of secondary structure content consistent with the design model were subjected to crystallographic analysis. One design based on Rossman 2x2 expressed as soluble protein, but no crystal structure

could be obtained. Crystal structures were obtained for two designed proteins: a ferredoxin-top7 chimera and a ferredoxin-ferredoxin chimera. The design and characterization of these two proteins is described in the Results.

Crystal structures were used to search for structural homologs in the SCOP database. First, crystal structures (ferredoxin-top7: 4KYZ chain A, ferredoxin-ferredoxin: 5CW9 chain A) were used as search queries using TMalign [24]. Hits were saved only if the alignment covered 75% or more of the query structure. Results were sorted by TM-score to identify the most similar structures in the SCOP database. Secondary structure topology cartoons were created with the Pro Origami server [25]. To map designed protein crystal structures into the protein domains network, the structures were aligned to all domain structures in the protein domains network using the PDBeFold server [26]. PDBeFold structural alignment hits were filtered for RMSD less than or equal to 2.5Å and aligned sequence length of greater than or equal to 75 residues. In contrast to the methods of Nepomnyachi et al, sequence similarity thresholds were ignored. Including sequence similarity thresholds eliminates matching hits in the domains network. This is not surprising because the proteins were designed de novo and did not evolve from natural proteins. Filtered alignment hits were mapped into the protein domains network using Cytoscape [27]. To evaluate neutral drift models of the parent folds, then crystal structures of de novo ferredoxin and Top7 proteins (2KL8 and 1QYS) were obtained and corresponding mutations from the final design proteins were modeled using a flexible backbone protein design algorithm described previously [23]. Final Rosetta energies were calculated and subtracted from the Rosetta energies of the original parent protein structures to obtain predictions of the change in free energy of folding.

The ferredoxin – TOP7 protein (NESF ID OR327) was expressed, and purified following standard protocols developed by the NESG for production of selenomethionine labeled protein samples [28]. Briefly, *Escherichia coli* BL21 (DE3) pMGK cells, a rare-codon enhanced strain, were transformed with the DNA sequence-verified OR327-21.1 plasmid. A single isolate was cultured in MJ9 minimal media supplemented with selenomethionine, lysine, phenylalanine, threonine, isoleucine, leucine, and valine for the production of selenomethionine-labeled OR327. Initial growth was carried out at 37 °C until the OD600 of the culture reached □0.8 units. The incubation temperature was then decreased to 17 °C, and protein expression was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) at a final concentration of 1 mM. Following overnight incubation at 17 °C, the cells were harvested by centrifugation and resuspended in Lysis Buffer [50 mM Tris, pH 7.5, 500 mM NaCl, 1 mM tris (2-carboxyethyl)phosphine, 40 mM imidazole]. After sonication, the supernatant was collected by centrifugation for 40 min at 30,000 × g. The supernatant was loaded first onto a Ni affinity column (HisTrap HP; GE Healthcare) and the eluate loaded into a gel filtration column (Superdex 75 26/60; GE Healthcare). Yields were 60-90 mg / L. The purified 6His-OR327 construct in buffer

containing 10 mM Tris·HCl, 100 mM NaCl, 5 mM DTT, pH 7.5, was then concentrated to □10.6 mg/mL The sample was flash-frozen in 50-μL aliquots using liquid nitrogen and stored at −80 °C before crystallization trials. The sample purity (>98%), molecular weight, and oligomerization state were verified by SDS/PAGE, MALDI-TOF mass spectrometry, and analytic gel filtration followed by static light scattering, respectively. For static light scattering, selenomethionine-labeled ferredoxin – TOP7 protein (30 μL at 10 mM Tris·HCl, pH 7.5, 100 mM NaCl, 5 mM DTT) was injected onto an analytical gel filtration column (Shodex KW-802.5; Shodex) with the effluent monitored by refractive index (Optilab rEX) and 90° static light-scattering (miniDAWN TREOS; Wyatt Technology) detectors.

ACCESSION CODES

Structures have been deposited in the Protein Data Bank as entries 5CW9, 4KYZ, and 4KY3.

AUTHOR INFORMATION

Corresponding Author

chrisk1@uw.edu

Present Addresses

University of Washington, Molecular Engineering and Sciences Building, 4th Floor, 3946 W Stevens Way NE, Box 351655, Seattle, WA 98195-1655

Competing Financial Interests

The authors declare no competing financial interests.

ACKNOWLEDGMENT

We thank Rie Koga and Nobuyasu Koga for data analyses and technical assistance. We thank Lei Mao for technical assistance. This work was supported by the Defense Threat Reduction Agency and by a grant from the National Institute of General Medical Sciences Protein Structure Initiative U54-GM094597 (to G.T.M., J.H.).

REFERENCES

1. Aroul-Selvam, R., Hubbard, T. & Sasidharan, R. Domain insertions in protein structures. *J Mol Biol* **338**, 633-41 (2004).

2. Berrondo, M., Ostermeier, M. & Gray, J.J. Structure prediction of domain insertion proteins from structures of individual domains. *Structure* **16**, 513-527 (2008).
3. Lupas, A.N., Ponting, C.P. & Russell, R.B. 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? *Journal of Structural Biology* **134**, 191-203 (2001).
4. Pandya, C. et al. Consequences of domain insertion on sequence-structure divergence in a superfold. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E3381-E3387 (2013).
5. Ay, J., Gotz, F., Borriss, R. & Heinemann, U. Structure and function of the Bacillus hybrid enzyme GluXyn-1: native-like jellyroll fold preserved after insertion of autonomous globular domain. *Proc Natl Acad Sci U S A* **95**, 6613-8 (1998).
6. Collinet, B. et al. Functionally accepted insertions of proteins within protein domains. *Journal of Biological Chemistry* **275**, 17428-17433 (2000).
7. Cutler, T.A., Mills, B.M., Lubin, D.J., Chong, L.T. & Loh, S.N. Effect of Interdomain Linker Length on an Antagonistic Folding-Unfolding Equilibrium between Two Protein Domains. *Journal of Molecular Biology* **386**, 854-868 (2009).
8. Doi, N. & Yanagawa, H. Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *Febs Letters* **453**, 305-307 (1999).
9. Edwards, W.R., Busse, K., Allemann, R.K. & Jones, D.D. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic acids research* **36**(2008).
10. Guntas, G. & Ostermeier, M. Creation of an allosteric enzyme by domain insertion. *Journal of Molecular Biology* **336**, 263-273 (2004).
11. Ostermeier, M. Engineering allosteric protein switches by domain insertion. *Protein Engineering Design & Selection* **18**, 359-364 (2005).
12. Grishin, N.V. Fold change in evolution of protein structures. *Journal of Structural Biology* **134**, 167-185 (2001).
13. Soding, J. & Lupas, A.N. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837-846 (2003).
14. Krishna, S.S. & Grishin, N.V. Structural drift: a possible path to protein fold change. *Bioinformatics* **21**, 1308-1310 (2005).
15. Friedberg, I. & Godzik, A. Connecting the protein structure universe by using sparse recurring fragments. *Structure* **13**, 1213-1224 (2005).
16. Ben-Tal, N. & Kolodny, R. Representation of the Protein Universe using Classifications, Maps, and Networks. *Israel Journal of Chemistry* **54**, 1286-1292 (2014).
17. Park, K. et al. Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* **22**, 167-74 (2015).
18. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222-7 (2012).
19. Hocker, B., Claren, J. & Sterner, R. Mimicking enzyme evolution by generating new (beta alpha)(8)-barrels from (beta alpha)(4)-half-barrels. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16448-16453 (2004).
20. Shanmugaratnam, S., Eisenbeis, S. & Hocker, B. A highly stable protein chimera built from fragments of different folds. *Protein Engineering Design & Selection* **25**, 699-703 (2012).
21. DiMaio, F. et al. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nature Methods* **10**, 1102-1104 (2013).
22. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* **247**, 536-540 (1995).
23. Huang, P.S. et al. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PloS one* **6**(2011).

24. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).
25. Stivala, A., Wybrow, M., Wirth, A., Whisstock, J.C. & Stuckey, P.J. Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics* **27**, 3315-3316 (2011).
26. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**, 2256-68 (2004).
27. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
28. Xiao, R. et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *Journal of Structural Biology* **172**, 21-33 (2010).

FIGURE SUPPLEMENT TITLES/CAPTIONS

Figure 1. Domain insertion strategy for combining ferredoxin-top7 (A) and ferredoxin-ferredoxin (B). Two beta strands from each partner (red and purple) are concatenated to form the central strand pair of the fusion protein (pink).

Figure 2. Crystal structure of ferredoxin-top7 (4KYZ, chain A) aligned with design model (A) showing core packing of the insert (B) and host (C) domains. Crystal structure colored by B-factor. Design model in gray.

Figure 2—figure supplement 1. Circular dichroism spectra showing alpha and beta structure at 25°C for ferredoxin-top7.

Figure 3. Crystal structure of ferredoxin-ferredoxin (5CW9) aligned with design model showing overall alignment of helices (A) and the fused beta sheet (B). Crystal structure colored by B-factor. Design model in gray.

Figure 3—figure supplement 1. Circular dichroism spectra showing alpha and beta structure at 25°C for ferredoxin-ferredoxin.

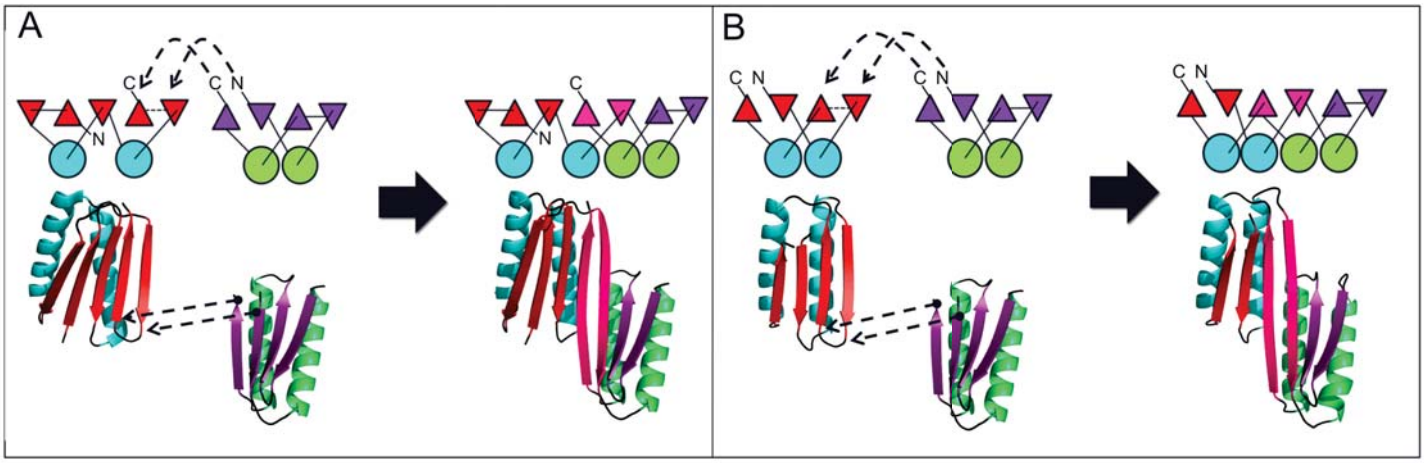
Figure 3—figure supplement 2. Ferredoxin-Ferredoxin 2Fo-Fc omit map superimposed with crystal structure shows core packing of host (A) and insert (B) domains.

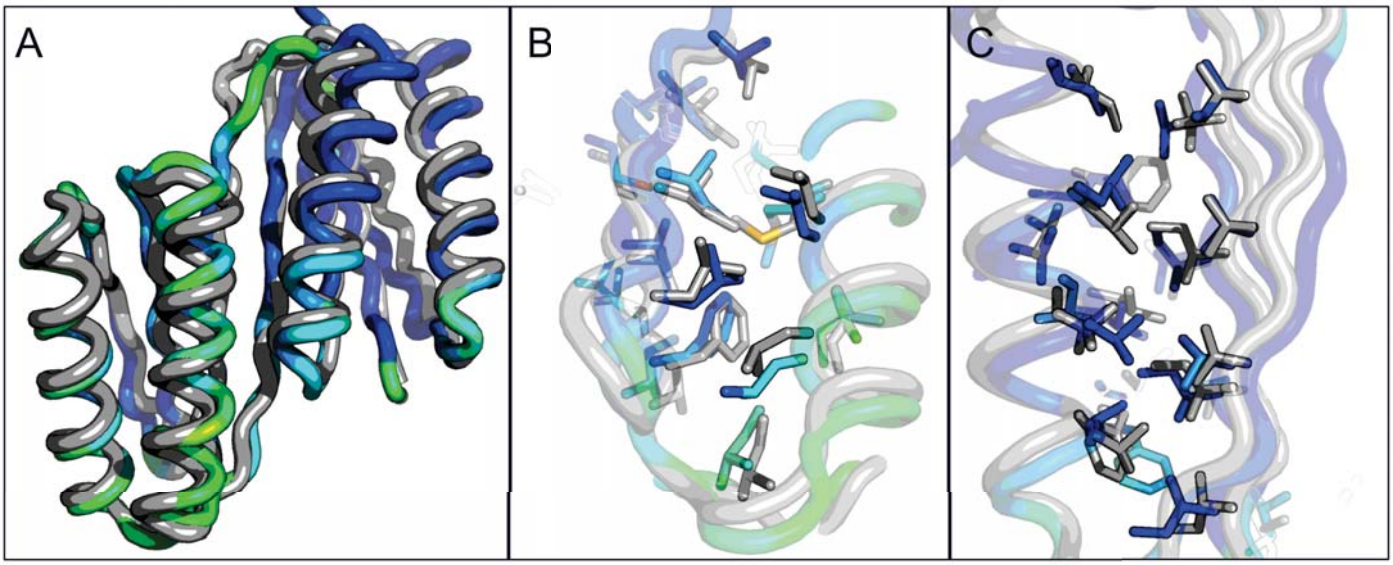
Figure 4. Top two SCOP domain structural homologues for Fd-Top7 (A) and Fd-Fd (B) designed domain as determined by TM-align scores.

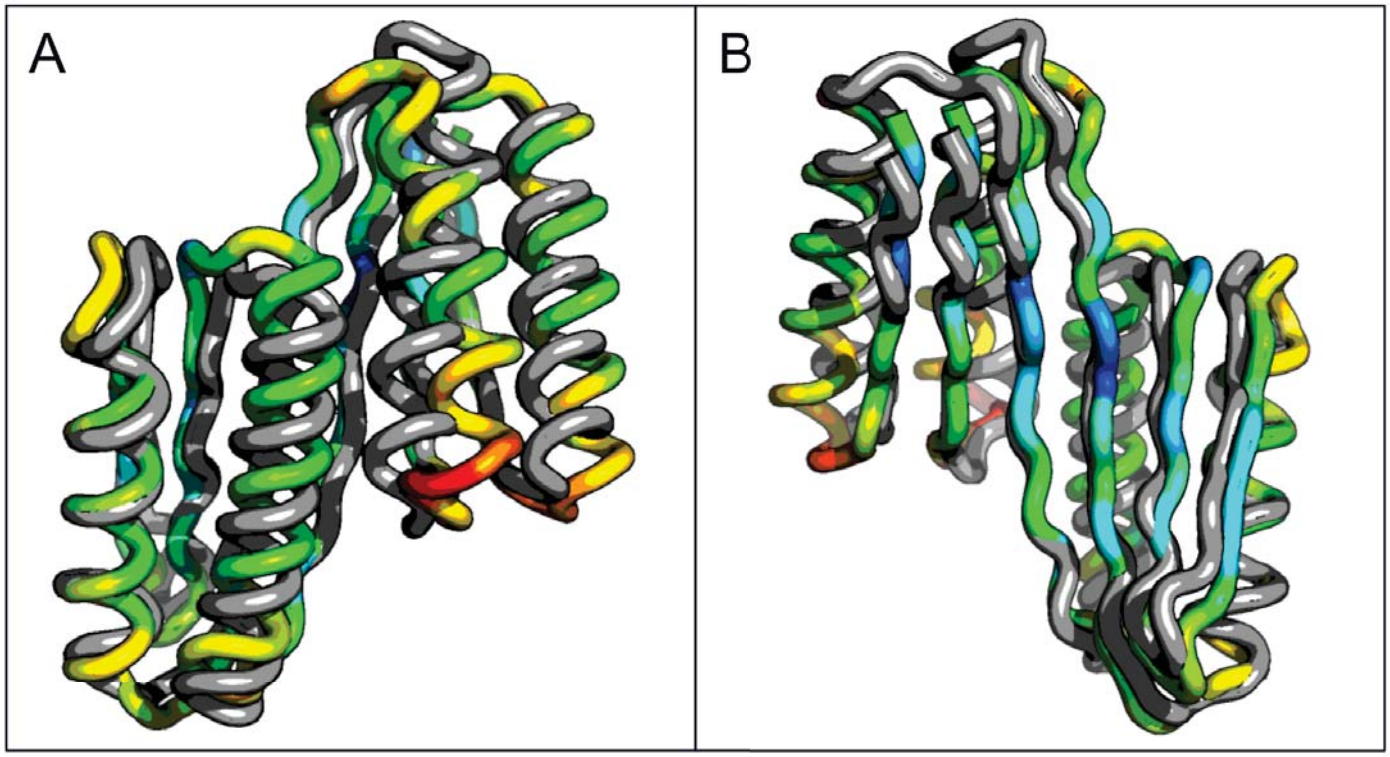
Figure 4—figure supplement 1. Parent domain PDB structures (2KL8, 1QYS) and daughter designed folds (5CW9,4KYZ) (pink) mapped into the α+β region of the SCOP domains network of Nepomnyachi et al. (A) and zoomed region (B) highlighting parent, designed, and first neighbor folds.

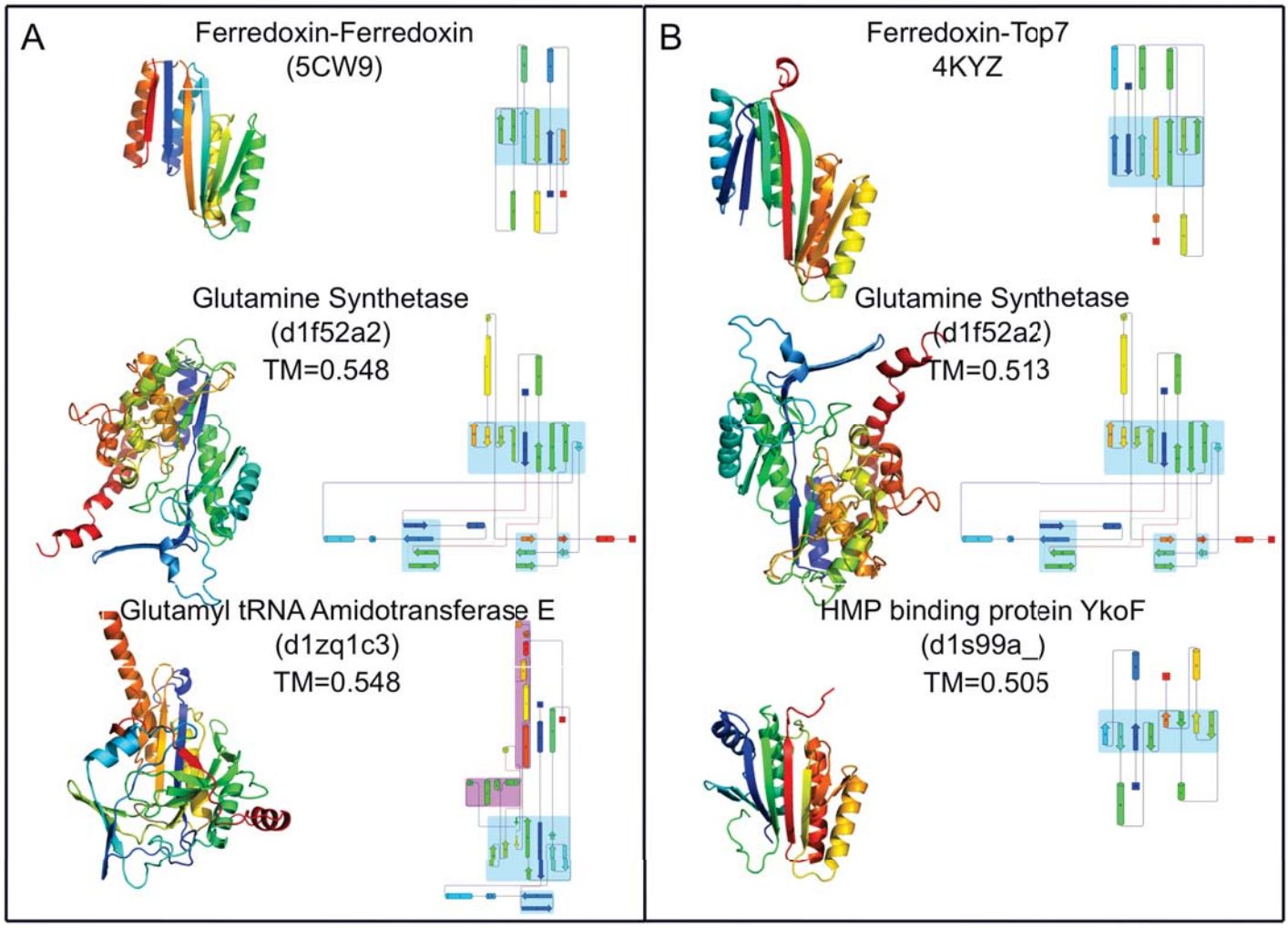
Figure 4—figure supplement 2. Neutral drift mutant models, relative changes to predicted free energy of folding in REU (Rosetta Energy Units), and multiple sequence alignment of parent and designed sequences, showing mutations in ferredoxin-top7 (A) and ferredoxin-ferredoxin (B).

Supplementary File 1. Crystallographic Data


A
C
C N
N
C
N
B
C N
C N
C N


A
B
C


A
B


A
Ferredoxin-Ferredoxin
(5CW9)
Glutamine Synthetase
(d1f52a2)
TM=0.548
Glutamyl tRNA Amidotransferase E
(d1zq1c3)
TM=0.548
B
Ferredoxin-Top7
4KYZ
Glutamine Synthetase
(d1f52a2)
TM=0.513
HMP binding protein YkoF
(d1s99a_)
TM=0.505