and NSF (MCB-1445201). Design calculations were facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. X-ray crystallography and SAXS data were collected at the Advanced Light Source (Lawrence Berkeley National Laboratory, Berkeley, California Department of Energy, contract no. DE-AC02-05CH11231); SAXS data were collected through the SIBYLS mail-in SAXS program under the aforementioned contract number, and we thank K. Burnett and G. Hura. The Berkeley Center for Structural Biology is supported in part by the National Institute of General Medical Sciences (NIH), and the Howard Hughes Medical Institute. G.O. is a Marie Curie International Outgoing Fellowship fellow (332094 ASR-CompEnzDes FP7- People-2012-IOF). B.G and J.M.G. are supported by Washington Research Foundation Innovation Postdoctoral Fellowships. Coordinates and structure files have been deposited to the Protein Data Bank with accession codes: 5J0J (2L6HC3_6), 5J0I (2L6HC3_12), 5J0H (2L6HC3_13), 5IZS (5L6HC3_1), 5J73 (2L4HC2_9), 5J2L (2L4HC2_11), 5J0L (3L6HC2_2), 5J0K (2L4HC2_23), 5J10 (2L4HC2_24). S.E.B., Z.C., and D.B. designed the research and S.E.B. and D.B. wrote the manuscript S.E.B. developed the HBNet method and wrote the program code. D.B. wrote the parametric backbone generation code with help from C.X. and G.O. A.F. wrote the loop closure program code. S.E.B., Z.C., R.A.L., and D.B. carried out design calculations. S.E.B. and Z.C. purified and biophysically characterized the designed proteins. B.G. performed yeast two-hybrid assays. J.M.G. performed mass spectrometry. J.H.P. crystallized the designed proteins. B.S. and P.H.Z. collected and analyzed crystallographic data. B.S., P.H.Z., G.O, and F.D. solved structures with help from S.E.B. and Z.C. All authors discussed results and commented on the manuscript.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/352/6286/680/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S20
Tables S1 to S5
References (*62–82*)

4 December 2015; accepted 23 March 2016
10.1126/science.aad8865

REPORTS

PROTEIN DESIGN

Design of structurally distinct proteins using strategies inspired by evolution

T. M. Jacobs,[1] B. Williams,[2] T. Williams,[2] X. Xu,[3,4]* A. Eletsky,[3,4] J. F. Federizon,[3] T. Szyperski,[3] B. Kuhlman[2,5]†

Natural recombination combines pieces of preexisting proteins to create new tertiary structures and functions. We describe a computational protocol, called SEWING, which is inspired by this process and builds new proteins from connected or disconnected pieces of existing structures. Helical proteins designed with SEWING contain structural features absent from other de novo designed proteins and, in some cases, remain folded at more than 100°C. High-resolution structures of the designed proteins CA01 and DA05R1 were solved by x-ray crystallography (2.2 angstrom resolution) and nuclear magnetic resonance, respectively, and there was excellent agreement with the design models. This method provides a new strategy to rapidly create large numbers of diverse and designable protein scaffolds.

Most efforts in de novo protein design have been focused on creating idealized proteins composed of canonical structural elements. Examples include the design of coiled coils, repeat proteins, TIM barrels, and Rossman folds (*1–6*). These studies elucidate the minimal determinants of protein structure, but they do not aggressively explore new regions of structure space. Additionally, idealized structures may not always be the most effective starting points for engineering novel protein functions. Functional sites in proteins are often created from nonideal structural elements, such as kinks, pockets, and bulges.

[1]Program in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [2]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [3]Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14260, USA. [4]Northeast Structural Genomics Consortium. [5]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
*Present address: Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA.
†Corresponding author. Email: bkuhlman@email.unc.edu

The lack of nonideal structural elements from de novo designed proteins highlights a key difference between natural protein evolution and current design methods. Specifically, protein design methods universally begin with a target structure in mind. Therefore, the space of designable structures that can accommodate these nonideal protein elements is limited by the imagination of the designer. In contrast, natural evolution is based not on design but on cellular fitness provided by the evolved protein function. This lack of a predetermined target fold is a powerful feature of protein evolution that holds significant potential for the design of novel structures and functions. In an effort to tap this potential, we sought to develop a method of computational protein design inspired by mechanisms of natural protein evolution.

Gene duplication and homologous recombination mix and match elements of protein structure to give rise to new structures and functions (*7–9*). This phenomenon is most evident at the level of independently folding protein domains (*10–12*), but recent studies have shown that these same principles function at a smaller scale during the evolution of distinct, globular protein folds (*13*). Insertions, deletions, and replacement of secondary and supersecondary structural elements sample alternative tertiary structures (*14–16*). Our design strategy, called SEWING (structure extension with native-substructure graphs), is motivated by this process and builds new protein structures from pieces of naturally occurring protein domains. The process is not dictated by the need to adopt a specific target fold but rather is aimed at creating large sets of alternative structures that satisfy predefined design requirements. One of the strengths of this approach is that it ensures that all of the structural elements of the protein are inherently designable, at the same time allowing for the incorporation of structural oddities unlikely to be found in idealized proteins. Here, we apply SEWING to the design of helical proteins. We show that designed structures are diverse and contain structural features absent from alternative design strategies.

SEWING begins with the extraction of small structural motifs, or substructures, from existing protein structures. These serve as the basic building blocks for all generated models. We aimed to identify substructures that were large enough to carry information regarding structural preference yet small enough to allow combinations that can generate novel globular structures. Ultimately, we chose to extract two distinct types of substructures. The first is composed of continuous stretches of protein structure that encompass two secondary structural elements separated by a loop (Fig. 1). These substructures capture the relative orientation between adjacent secondary structure elements and maintain local packing interactions. In addition, there is evidence that substructures of this size adopt a relatively limited number of conformations that have already been sampled exhaustively in known protein structures (*14*). The second type is composed of groups of three to five secondary structural elements, where each element makes van der Waals contacts with every other element, but the elements are not necessarily continuous in primary sequence (Fig. 1, supplementary methods). Nonadjacent, or discontinuous, substructures maintain longer-range tertiary interactions that provide valuable stability and are often conserved during protein evolution (*17*).

The goal of SEWING is to combine and modify these extracted components in order to develop new tertiary structures. Naturally occurring homologous recombination, in which sequence similarity between DNA molecules leads to the combination of the genetic material, guides the formation of new protein chimeras. This process enriches for proteins that are more likely to be well folded and functional, as sequence-similarity filters for segments that are structurally compatible. In the case of SEWING, we know the three-dimensional structures of the building blocks; therefore, we can directly use structural information to probe which substructures are well suited for combination. During SEWING, continuous substructures are eligible for combination if the C-terminal region of one substructure shares high structural similarity with the N-terminal region of another substructure and if superposition of the two regions does not create any steric clashes between other regions in the two substructures. This type of superposition ensures that the three-dimensional spacing between all pairs of secondary structural elements adjacent in primary sequence is similar to that observed in the Protein Data Bank (PDB). During discontinuous SEWING, combinations are created by superimposing two elements (helices in this study) from one substructure with two elements from another substructure. For both continuous and discontinuous SEWING, structure similarity is identified by using a fast geometric hashing approach that ensures that the regions of interest can align with low root mean square deviation (RMSD) (*18*).

Once pairwise structural similarity is calculated between all substructures, these data are used to generate a large graph (Fig. 1). The nodes in this graph represent the substructures, and the edges indicate a level of structural similarity that allows recombination. Novel structures are generated from this graph by traversing a path wherein each followed edge adds new structural elements to the design model. The number of edges included in the sampled paths can control the approximate size of the generated structures. Unlike previously described methods of de novo backbone generation, no target structure is required, and output structures span a diverse set of globular folds.

Previous studies have demonstrated that protein fragments can adopt alternative structures when placed in new environments (*19*–*21*), and thus, similar to natural evolution, the next step in the design process was to further stabilize the protein through mutagenesis. This optimization step was achieved using iterative steps of side-chain packing and backbone minimization available in the Rosetta molecular-modeling suite (*22*). Preference for the amino acid sequence present in the parental substructure was used to better preserve the structural interactions inherent to the parent substructures.

To test SEWING, we designed a diverse set of helical proteins using graphs composed of continuous substructures or discontinuous substructures. Continuous and discontinuous substructures were extracted from nonredundant subsets of the PDB (*23*, *24*). In total, 33,928 continuous substructures and 4584 discontinuous substructures were extracted. Design models from the continuous graph were generated by using three-edge paths and were therefore composed of substructures extracted from four existing structures from the PDB (Fig. 1). The continuous graph contained 345 million edges, which allowed an estimated 7×10^{16} backbones that can be filtered and optimized in later designs steps (supplementary methods). Initially, 50,000 alternative tertiary structures were created and used as templates for rotamer-based sequence optimization and energy minimization. These models were filtered and sorted by using metrics that evaluate predicted energy (normalized by sequence length), side-chain packing, buried polar groups, and sequence-structure agreement (supplementary methods) (*25*). When examining the models, we noticed that the naïve SEWING procedure was biased toward creating low–contact order models, i.e., structures with few contacts between residues distant in primary sequence. To overcome this bias, we filtered for models with contact orders more representative of naturally occurring helical proteins (fig. S1). We have subsequently demonstrated that Monte Carlo sampling of the SEWING graph with a score function that favors long-range contacts can be used to build high-contact models with high frequency (fig. S1). This illustrates one way

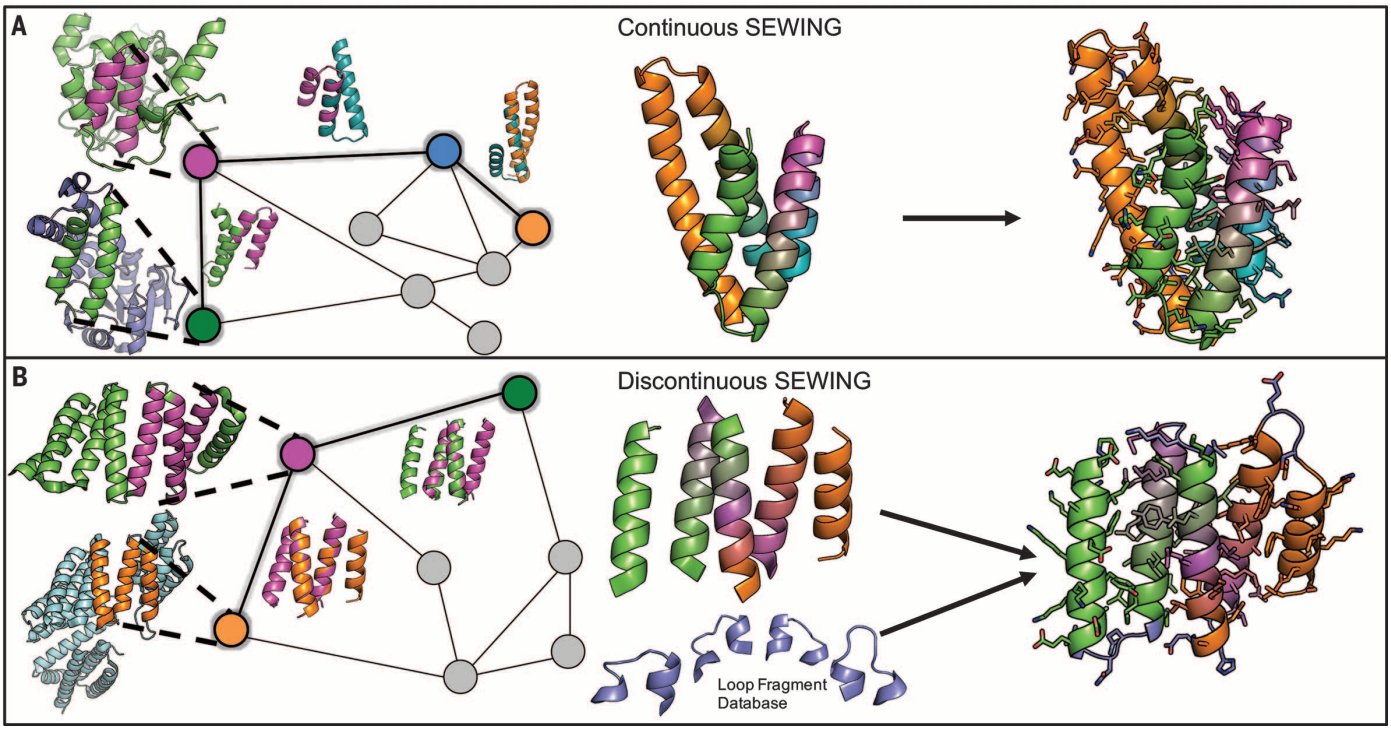

Fig. 1. Overview of the SEWING method. (**A**) Continuous SEWING workflow for CA01. (**B**) Discontinuous SEWING workflow for DA05. From left to right: Parental PDBs are shown with extracted substructures; graph schematic—colored nodes indicate substructures contained in the final design model, and superimposed structures show structural similarity indicated by adjacent edges; design model before sequence optimization and loop design; and final design models.

that directed sampling of the SEWING graph can be used to enforce design requirements.

In total, 11 designs based on continuous SEWING were selected for experimental characterization (table S1). Each region of the final designs shared between 45 and 65% sequence identity with the substructure that they were built from (figs. S2 and S3), but when performing a BLAST search with the full-length sequences, no matches were identified that aligned over the full length of the proteins. Eight designs expressed well in *Escherichia coli* and were readily purified from the soluble fraction of 1-liter cultures. Four of the eight proteins were monomeric as seen by size-exclusion chromatography–multiangle light scattering, had a circular dichroism (CD) spectrum characteristic of a helical protein, and unfolded cooperatively upon thermal denaturation (Fig. 2 and figs. S4 and S5). Two of the designed proteins are hyperthermophiles and require high concentrations of chemical denaturant in order for one to observe thermal unfolding (Fig. 2B). For one design, CA01, several thermodynamic parameters were determined by fitting a modified Gibbs-Helmholtz equation to the thermal and chemical denaturation surface (Fig. 3B and table S2) (*26*). The extrapolated melting temperature of 126°C places it among the top 0.01% of values in the ProTherm database of protein stabilities (*27*). The crystal structure of CA01 was solved to 2.2 Å and shows excellent agreement with the design model, with the RMSD of the *Cα* atomic coordinates equaling 0.8 Å. Similarly, the side-chain packing of the protein core is nearly identical for the design model and the experimental structure (Fig. 3, fig. S6, and table S3).

The structural variety in the design models for the well-folded proteins is of particular note (Fig. 2). The SEWING-generated models include kinked and curved helices (figs. S7 and S8), cavities and clefts (figs. S9 and S10), and a large range of helix-crossing angles (Fig. 2). The topologies of the SEWING models are varied when compared with previously designed α-helical proteins, which are restricted to coiled coils, repeat proteins, and up-down four-helical bundles (Fig. 2C). To compare SEWING models with naturally occurring protein structures, we searched for structurally similar domains using the Dali server (*28*). In general, large portions of the models aligned to regions of existing protein structures. However, the sequence identities across the alignments were typically below 20%, and the positions of the unaligned residues frequently diverged (fig. S11). For instance, the fifth helix of CA01 is shifted by ~9 Å relative to the fifth helix in the top Dali match. These sequences and structural differences provide unique surfaces that may serve as templates for future design goals.

To test discontinuous SEWING, models were generated from two-edge paths and, thus, were composed of structural elements from three parent structures. The variable number of helical elements in the discontinuous substructures therefore allowed design models to be composed of 5 to

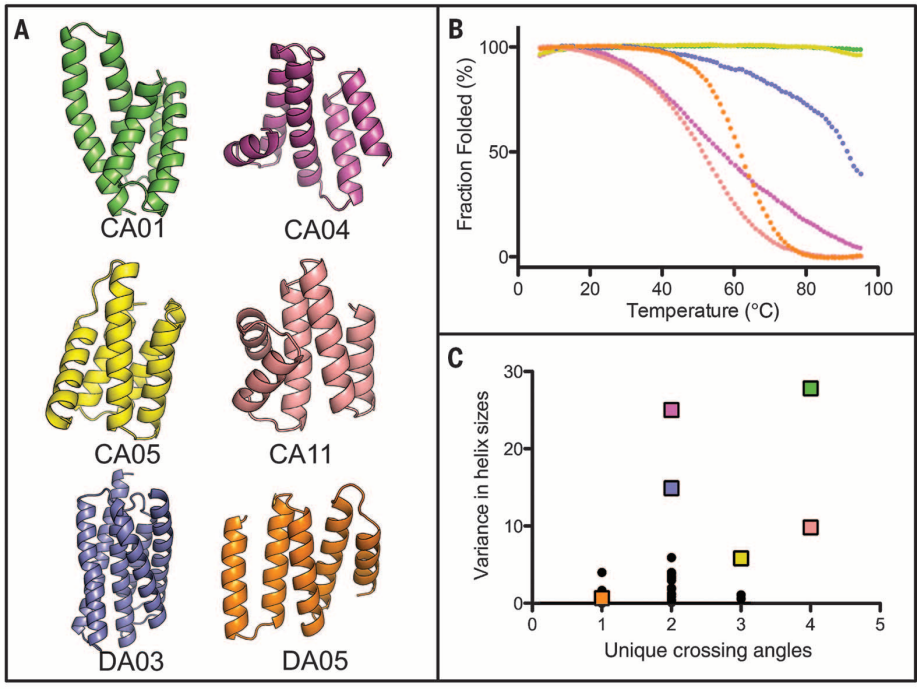

Fig. 2. Well-folded SEWING designs. (**A**) Design models obtained with continuous (CA) and discontinuous (DA) SEWING. (**B**) Temperature denaturation curves for well-folded SEWING designs, colored to match design models. (**C**) A comparison of previously designed helical structures (black dots) to SEWING models (colored squares) demonstrates the structural complexity of SEWING designs. We calculated crossing angles between all pairs of helices in each structure. Crossing angles were considered unique if they differed by >20 degrees from all other calculated angles in the same structure. Variance in helix size describes the calculated variance in the number of residues per helix for all helices in a single structure. A complete list of helix and crossing-angle definitions for de novo designs can be found in tables S5 and S6.

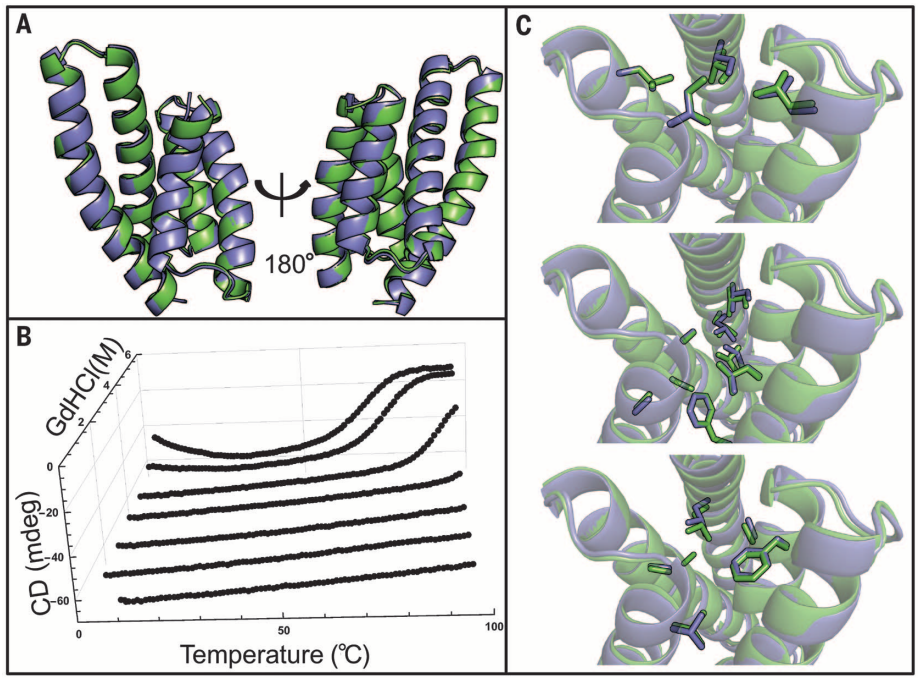

Fig. 3. Structural and biophysical characterization of CA01. (**A**) Backbone superimposition of the design model (green) and crystal structure (purple). (**B**) In the chemical and temperature denaturation experiment, a sharp unfolding transition is observed at 5 M GdHCl and 75°C. (**C**) Comparison of side-chain packing between the design model (green) and crystal structure (purple) at three different layers of the structure.

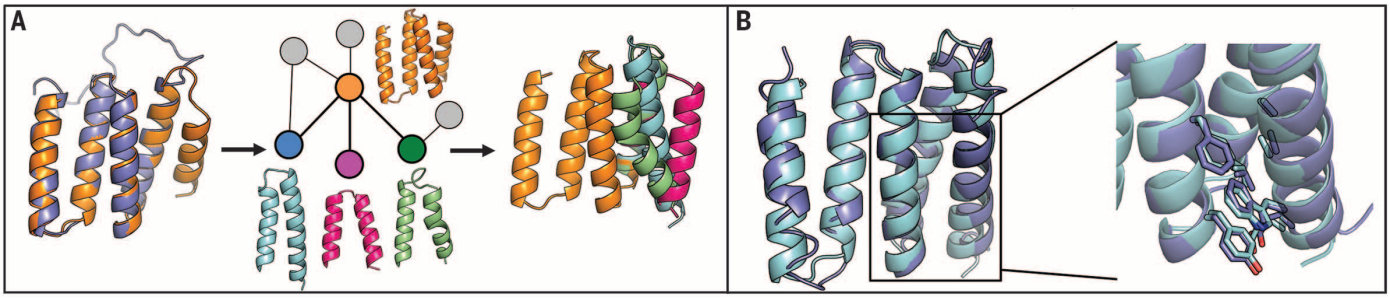

Fig. 4. Result for discontinuous assembly DA05 and DA05R1. (**A**) From left to right: Backbone superimposition of the DA05 design model (orange) with a member of the NMR ensemble (blue). An example of continuous substructure graph for the design of a new final helix onto DA05. Superimposition of three design models containing new helices. (**B**) From left to right: Backbone superimposition of the DA05R1 design model (light blue) with a member of the DA05R1 NMR ensemble (blue). A comparison of side-chain packing between the DA05R1 design model and the NMR structure for DA05R1.

11 helices. Unlike models from the continuous-substructure graph, discontinuous models require the addition of loops between consecutive helices. Loops were designed by using a database of fragments from the PDB (*29*). Each loop fragment was superimposed onto the design model and optimized using gradient-based minimization in Cartesian space. Any path that created structures for which no loop fragment could be found was eliminated from the set of designs. Design models were filtered and optimized in the same way as models from the continuous graph. In total, 10 were selected for experimental characterization (table S1).

Of these 10 designs, 2 expressed at levels sufficient for purification. Both purified proteins were helical and folded, as evidenced by CD (Fig. 2 and fig. S4). Similar to the results from the continuous designs, one discontinuous design, DA03, demonstrated high thermostability, which required high levels of denaturant to completely unfold. For this design, an 181-residue six-helix bundle, unfolding appears to follow a three-state model (fig. S12).

The structure of the other well-folded discontinuous design, DA05, was solved using nuclear magnetic resonance (NMR) spectroscopy, as the protein did not readily crystallize (figs. S13 and S14 and table S4). The first four helices of the design model match the lowest-energy member of the NMR ensemble very closely, with a Cα RMSD of 0.8 Å (Fig. 4 and fig. S6). However, the NMR data indicate that the final helix of the protein is disordered in solution. In an effort to identify the errors in the design model that led to the unstructured region, structural preference for the designed sequence was evaluated with fragment analysis as described previously (*1*). The fragments extracted for the unstructured region showed especially poor preference for the designed helical structure (fig. S15). We attempted to design a new final helix for the DA05 design using the continuous SEWING method. The final helix of the initial design model was removed, and the remainder of the model was added as a node to the continuous graph. New helices were evaluated by following a single edge from this new node (Fig. 4A). Three models designed in this way were selected for experimental testing. Two of the tested designs, DA05R1 and DA05R2, show a significant increase in melting temperature relative to the initial DA05 design (fig. S16). The NMR structure of DA05R1 shows that the newly designed helix adopts the designed conformation, which highlights the utility of combining the continuous and discontinuous graphs (Fig. 4B, figs. S17 and S18, and table S4).

The additional step of loop building is a critical difference between discontinuous and continuous SEWING. The accurate design of loops is a long-standing challenge for protein design, and this additional step may have contributed to the relatively lower success rate observed for discontinuous SEWING. In contrast, continuous SEWING maintains the relative orientation between adjacent helices, which allows many of the designed loop sequences to be taken directly from the native substructure. The power of this strategy is seen in the high structural accuracy achieved for the loops in the CA01 design (Fig. 3 and fig. S2).

Our results show that computational adaptations of basic evolutionary principles, such as recombination and mutation, can be used to design, accurately and rapidly, a diverse set of helical protein structures. The diversity of SEWING designs will further increase when alternative types of substructures are included, such as β-α motifs and β hairpins. Furthermore, discontinuous and continuous SEWING can be merged, as in the DA05R1 design, to create additional diversity. We anticipate that this structural diversity will be advantageous for functional design, as every backbone generated with SEWING has new surface and pocket features that provide potential binding sites for ligands or macromolecules. Additionally, SEWING offers an approach for stitching together functional motifs from naturally occurring proteins, an evolutionary mechanism to generate multifunctional proteins and allosteric systems.

REFERENCES AND NOTES

1. N. Koga *et al.*, *Nature* **491**, 222–227 (2012).
2. N. H. Joh *et al.*, *Science* **346**, 1520–1524 (2014).
3. P.-S. Huang *et al.*, *Science* **346**, 481–485 (2014).
4. L. Doyle *et al.*, *Nature* **528**, 585–588 (2015).
5. T. J. Brunette *et al.*, *Nature* **528**, 580–584 (2015).
6. B. Kuhlman *et al.*, *Science* **302**, 1364–1368 (2003).
7. A. L. Hughes, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8791–8792 (2005).
8. C. C. F. Blake, *Nature* **273**, 267–267 (1978).
9. M. Bashton, C. Chothia, *Structure* **15**, 85–99 (2007).
10. S. Koide, *Curr. Opin. Biotechnol.* **20**, 398–404 (2009).
11. S. Eisenbeis *et al.*, *J. Am. Chem. Soc.* **134**, 4019–4022 (2012).
12. C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, S. A. Teichmann, *Curr. Opin. Struct. Biol.* **14**, 208–216 (2004).
13. N. V. Grishin, *J. Struct. Biol.* **134**, 167–185 (2001).
14. N. Fernandez-Fuentes, J. M. Dybas, A. Fiser, *PLOS Comput. Biol.* **6**, e1000750 (2010).
15. J. Söding, A. N. Lupas, *BioEssays* **25**, 837–846 (2003).
16. G. A. Reeves, T. J. Dallman, O. C. Redfern, A. Akpor, C. A. Orengo, *J. Mol. Biol.* **360**, 725–741 (2006).
17. H. E. Aronson, W. E. Royer Jr., W. A. Hendrickson, *Protein Sci.* **3**, 1706–1711 (1994).
18. R. Nussinov, H. J. Wolfson, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 10495–10499 (1991).
19. M. J. Schellenberg *et al.*, *J. Mol. Biol.* **402**, 720–730 (2010).
20. T. A. M. Bharat, S. Eisenbeis, K. Zeth, B. Höcker, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9942–9947 (2008).
21. S. de Bono, L. Riechmann, E. Girard, R. L. Williams, G. Winter, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1396–1401 (2005).
22. A. Leaver-Fay *et al.*, *Methods Enzymol.* **487**, 545–574 (2011).
23. G. Wang, R. L. Dunbrack Jr., *Bioinformatics* **19**, 1589–1591 (2003).
24. J. S. Richardson, D. C. Richardson, *Biopolymers* **99**, 170–182 (2013).
25. W. Sheffler, D. Baker, *Protein Sci.* **19**, 1991–1995 (2010).
26. B. Kuhlman, D. P. Raleigh, *Protein Sci.* **7**, 2405–2412 (1998).
27. A. Sarai *et al.*, *Biopolymers* **61**, 121–126 (2001–2002).
28. L. Holm, P. Rosenström, *Nucleic Acids Res.* **38** (Web Server), W545–W549 (2010).
29. M. D. Tyka, K. Jung, D. Baker, *J. Comput. Chem.* **33**, 2483–2491 (2012).

ACKNOWLEDGMENTS

This work was supported by NIH grants R01GM073960 and R01GM117968 (to B.K.) and GM094597 (to T.S.). Use of the Advanced Photon Source was supported by the Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy, under contract no. W-31-109-Eng-38. Coordinates and structure factors have been deposited in the Protein Data Bank with the accession codes 5E6G (CA01), 2N8I (DA05), and 2N8W (DA05R1). Chemical shifts have been deposited in the Biological Magnetic Resonance Bank (BMRB) with the accession codes 25850 (DA05) and 25868 (DA05R1). T.M.J. and B.K. designed the research; T.M.J. wrote the backbone assembly code; T.M.J. and B.W. carried out the backbone assembly and design simulations; T.M.J. conducted the biophysical analysis; T.W. and T.M.J. solved the structure of CA01; and X.X., A.E., and J.F.F., with advice from T.S., solved the NMR structure of DA05 and DA05R1.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/352/6286/687/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S18
Tables S1 to S6
References (*30*–*55*)

5 November 2015; accepted 14 March 2016
10.1126/science.aad8036

Editor's Summary

Building new proteins from the old

Proteins are the workhorses of biology. Designing new, stable proteins with functions desirable in biotechnology or biomedicine remains challenging. Jacobs *et al.* developed a computational method called SEWING that designs proteins using pieces of existing structures (see the Perspective by Netzer and Fleishman). The new proteins can contain structural features such as pockets or grooves that are required for function. The solved structures of two designed proteins agreed well with the design models. The method allows rapid design of a diverse set of structures that will facilitate functional design.

Science, this issue p. 687; see also p. 657