Global analysis of protein folding using massively parallel design, synthesis, and testing

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevaller, Cheryl H. Arrowsmith, David Baker

Proteins fold into unique native structures stabilized by thousands of weak interactions that collectively overcome the entropic cost of folding. Although these forces are “encoded” in the thousands of known protein structures, “decoding” them is challenging because of the complexity of natural proteins that have evolved for function, not stability. We combined computational protein design, next-generation gene synthesis, and next-generation sequencing to systematically examine how sequence determines folding and stability in uncharted protein space. Iteration between design and experiment increased the design success rate from 6% to 74%, produced stable proteins unlike those found in nature for topologies where design was initially unsuccessful, and revealed subtle contributions to stability as designs became increasingly optimized. Our approach achieves the long-standing goal of accurately and efficiently model the balance among the many energy terms that contribute to the free energy of folding (1–3). Minimal protein domains (30 to 50 amino acids in length), such as the villin headpiece and WW domain, are commonly used to investigate this balance because they are the simplest protein folds found in nature (4). The primary experimental approach used to investigate this balance has been mutagenesis (5–12), but the results are context-dependent and do not provide a global view of the contributions to stability. Molecular dynamics simulations on minimal proteins have also been used to study folding (13–15), but these do not reveal which interactions specify and stabilize the native structures.

De novo protein design has the potential to reveal the sequence determinants of folding for minimal proteins by charting the space of nonnatural sequences and structures to define what can and cannot fold. Protein sequence space (16) is vastly larger than the set of natural proteins that currently form the basis for nearly all models of protein stability (9, 12, 17–19) and is unbiased by selection for biological function. However; only two minimal proteins (<50 amino acids, stabilized exclusively by noncovalent interactions) have been computationally designed to date: FSD-1 (20) and DS119 (21). In part, this is attributable to the cost of gene synthesis, which has limited such studies to testing tens of designs at most—a minuscule fraction of design space. Because of the small sample sizes, design experiments are typically unable to determine why some designs are stable and others are unstructured, resemble molten globules, or form aggregates (22).

Here, we present a new synthetic approach to examining the determinants of protein folding by exploring the space of potential minimal proteins using de novo computational protein design, with data generated by parallel DNA synthesis and protein stability measurements. To encode our designs, we use oligo library synthesis technology (23, 24), which was originally developed for transcriptional profiling and large gene assembly applications and is now capable of parallel synthesis of 10^4 to 10^5 arbitrarily specified DNA sequences long enough to encode short proteins (fig. S1). To assay designs for stability, we express these libraries in yeast so that every cell displays many copies of one protein sequence on its surface, genetically fused to an expression tag that can be fluorescently labeled (25) (Fig. 1A). Cells are then incubated with varying concentrations of protease, those displaying resistant proteins are isolated by fluorescence-activated cell sorting (Fig. 1B), and the frequencies of each protein at each protease concentration are determined by deep sequencing (Fig. 1C; for reproducibility of the assay, see fig. S2). We then infer protease EC_{50} values (the protease concentration at which one-half of the cells pass the collection threshold) for each sequence from these data by modeling the complete selection procedure (Fig. 1D) (26). Finally, each design is assigned a “stability score” (Fig. 1E): the difference between the measured EC_{50} (on a log_{10} scale) and the predicted EC_{50} in the unfolded state, according to a sequence-based model parameterized using EC_{50} measurements of scrambled sequences (figs. S3 and S4). A stability score of 1 corresponds to an EC_{50} value that exceeds the predicted EC_{50} in the unfolded state by a factor of 10. The complete experimental procedure costs less than $7000 in reagents (mainly from DNA synthesis and sequencing) and requires ~10 hours of sorting per protease for each library.

Massively parallel measurement of folding stability

Proteolysis assays have been used to select for stable sequences (27–29) and to quantify stability for individual proteins (30) and proteins from cellular proteomes (31), but to date they have not been used to quantify stability for all sequences in a constructed library. To evaluate the ability of the assay to measure stability on a large scale, we obtained a synthetic DNA library encoding four small proteins (Pn1 WW domain (32), HAYP65 WW domain (5, 10), villin HP35 (7, 11), and BBL (8)) and 116 mutants of these proteins whose stability has been characterized in experiments on purified material. The library also contained 19,610 unrelated sequences (a fourth-generation designed protein library; see below), and all sequences were assayed for stability simultaneously. Although the stability score is not a direct analog of a thermodynamic parameter; stability scores measured with trypsin and separately measured with chymotrypsin were each well correlated with folding free energies (or melting temperatures) for all four sets of mutants, with r^2 values ranging from 0.63 to 0.85 (Fig. 1, F to I). Most mutants in this data set were predicted to have unfolded-state EC_{50} values similar to those of their parent sequences, so the relative stability scores of the mutants are very similar to their relative EC_{50} values. However, in the case of villin assayed with chymotrypsin, the unfolded-state model improved the correlation between protease resistance and folding free energy from r^2 = 0.46 (using raw EC_{50} values) to the reported r^2 = 0.77 by correcting for the effect of mutations such as Lys^{50} → Met and Phe^{51} → Leu on intrinsic chymotrypsin cleavage rates. The mutual agreement among trypsin results, chymotrypsin results,
and experiments on purified protein indicates that the assay provides a robust measure of folding stability for small proteins.

**Massively parallel testing of designed miniproteins**

We selected four protein topologies (ααα, βββ, αββ, and ββα) as design targets. These topologies have increasing complexity: The ααα topology features only two loops and exclusively local secondary structure (helices); the βββ topology features only two loops and exclusively local secondary structure (helices); the αββ topology features four loops and features a mixed (parallel/antiparallel β sheet bridging the N and C termini). Of these topologies, only ααα proteins have been found in nature within the target size range of 40 to 43 residues; no proteins have been previously designed in any of the four topologies at this size (excluding designed ααα and βββ proteins stabilized by multiple disulfide linkages (33)). For each topology, we first designed between 5000 and 40,000 de novo proteins using a blueprint-based approach described in (34). Each design has a unique three-dimensional main-chain conformation and a unique sequence predicted to be near-optimal for that conformation. We then selected 1000 designs per topology for experimental testing by ranking the designs by a weighted sum of their computed energies and additional filtering terms (26). The median sequence identity between any pair of tested designs of the same topology ranged from 15% to 35%, and designs were typically no more than 40 to 65% identical to any other design. This diversity is due to the different backbone conformations possible within a topology, along with the vast sequence space available even for small proteins (fig. S5). For each design, we also included two control sequences in our library: one made by scrambling the order of amino acids in that design (preserving the overall amino acid composition), and a second made by scrambling the order while preserving both the composition and the hydrophobic or polar character at each position (35–37). The library comprised 12,459 different sequences in total: 4153 designed proteins and 8306 control sequences. The designed proteins are named according to their secondary structure topology (using H for α helix and E for β strand), their design round, and a design number.

We assayed the sequence library for stability using both chymotrypsin and trypsin. To stringently identify stable designs, we ranked sequences by the lower of their trypsin or chymotrypsin stability scores, referred to simply as their (overall) stability score from here on. The fully scrambled sequences and patterned scrambled sequences had similar stability score distributions; most

---

**Fig. 1. Yeast display enables massively parallel measurement of protein stability.**

(A) Each yeast cell displays many copies of one test protein fused to Aga2. The C-terminal c-Myc tag is labeled with a fluorescent antibody. Protease cleavage of the test protein (or other cleavage) leads to loss of the tag and loss of fluorescence. FITC, fluorescein isothiocyanate. (B) Libraries of 10^4 unique sequences are sorted by flow cytometry. Most cells show high protein expression (measured by fluorescence) before proteolysis (blue). Only some cells retain fluorescence after proteolysis; those above a threshold (shaded green region) are collected for deep sequencing analysis. (C) Sequential sorting at increasing protease concentrations separates proteins by stability. Each sequence in a library of 19,726 proteins is shown as a gray line tracking its change in population fraction relative to that in the preselection library (enrichment). Enrichment traces for seven proteins at different stability levels are highlighted in color. (D) EC50 values for the seven highlighted proteins in (C) are plotted on top of the overall density of the 46,187 highest-confidence EC50 measurements from design rounds 1 to 4. (E) Same data as at left, showing that stability scores (EC50 values corrected for intrinsic proteolysis rates) correlate better than raw EC50 values between the proteases. (F to I) Stability scores measured in high-throughput correlate with individual folding stability measurements for mutants of four small proteins. The wild-type sequence in each set is highlighted as a red circle. Credible intervals for all EC50 measurements are provided in (26). (F) Pin1 ΔG_unf data at 40°C from (32) by thermal denaturation. (G) hYAP65 melting temperature (Tm) data from (5, 10). (H) Villin HP35 ΔG_unf data at 25°C from (7, 11) by urea denaturation. (I) BBL ΔG_unf data at 10°C from (8) by thermal denaturation.
Fig. 2. Iterative, high-throughput computational design generates thousands of stable proteins and reveals stability determinants.

(A) Stability data for designs and control sequences, separated by topology (ααα, βαββ, ββαα, and ββββ) and by design round (1 to 4). For each round and topology, the upper plot shows the total number of designed proteins (y axis) exceeding a given stability score threshold (x axis; stability increases from left to right). The number of designs tested (top left) may be lower than the number originally ordered (described in the text) because low-confidence data were removed (gray), and hydrophobic-polar pattern designed sequences (colors), fully scrambled sequences (“Scramb,” light gray), and hydrophobic-polar pattern–preserving scrambled sequences (“Pattern,” dark gray). Categories in rounds 2 to 4 were designs, patterned scrambles, and point mutants of designs, with single Asp mutations expected to be destabilizing (“BuryAsp,” yellow). (B to G) Determinants of stability from rounds 1 to 3 [as labeled in (A)]. Colored histograms show the number of tested designs (left y axis) in each bin for the structural metric on the x axis. Black lines show the success rate (fraction of designs tested with stability score > 1.0, right y axis) within a moving window the size of the histogram bin width, with a shaded 95% confidence interval from bootstrapping. Design success is shown as a function of NPSA from hydrophobic residues [(B), (D), (E), and (F)]; as a function of geometric agreement between 9-residue fragments of similar sequences in the design models and natural proteins [see text and (26)], measured in average RMSD (C); and as a function of Rosetta total energy (G).

(H) Overall success rate and number of successful designs per round (stability score > 1.0 with both proteases) for all topologies across all rounds. (I) Design success as a function of predicted success according to the topology-specific logistic regression models used to select round-4 designs for testing (trained on data from rounds 1 to 3). As in (B) to (G), colored histograms indicate the number of tested designs at each level of predicted success (left y axis), and the black line indicates the success rate (right y axis). See fig. S8 for individual success rates for each topology.

To examine this further, we selected six stable designs (four ααα and two βαββ) for Escherichia coli expression, purification, and further characterization by size exclusion chromatography (SEC) and circular dichroism (CD) spectroscopy. All six designs eluted from SEC as expected for a 5- to 7kDa monomer, and the CD spectra were consistent with the designed secondary structure (fig. S6A and table S1). Five of the six designs had clear, cooperative melting transitions, refolded reversibly, and were highly stable for minimal proteins: All had melting temperatures above 70°C, and the βαββ design EHEE_rd1_0284 had only partially melted at 95°C (free energy of unfolding ΔGunf = 4.7 kcal/mol; Fig. 3D). The sixth design, HHH_rd1_0005, did not refold and showed signs of aggregation (fig. S6A). We determined solution structures for EHEE_rd1_0284 and the left-handed ααα design HHH_rd1_0142 by nuclear magnetic resonance (NMR); each structure closely matched the design model [average backbone root mean square deviation (RMSD) = 2.2 Å for each NMR ensemble member against the design model] (Fig. 3A; see table S2 for NMR data summary). In sum, both high-throughput control experiments and
low-throughput characterization of individual proteins indicate that the protease-resistant designs folded as designed.

**Global determinants of stability**

This large set of stable and unstable minimal proteins with varying physical properties enabled us to quantitatively examine which protein features correlated with folding. We computed more than 60 structural and sequence-based metrics and examined which metrics differed between the 195 most stable αααα designs (stability score > 1.0, considered to be design successes) and the 664 remaining αααα designs (considered to be failures) using the Kolmogorov-Smirnov two-sample test. Significant differences indicate that a particular metric captures an important contribution to protein stability and that this contribution was poorly optimized among the tested designs.

The dominant difference between stable and unstable αααα designs was the total amount of buried nonpolar surface area (NPSA) from hydrophobic amino acids (Fig. 2B). Stable designs buried more NPSA than did unstable designs ($P < 5 \times 10^{-38}$; fig. S7A), and none of the 95 designs below 32 Å² per residue were stable. Above this threshold, the success rate (ratio of successful designs to tested designs) steadily increased as buried NPSA increased (Fig. 2B). Stable designs also had better agreement between their sequences and their local structures, as assessed by quantifying the geometric similarity (in Å of RMSD) between 9-residue fragments of the designs and 9-residue fragments of natural proteins similar in local sequence to the designed fragment (Fig. 2C) (26). Fragments of stable designs were more geometrically similar to fragments of natural proteins of similar local sequence, whereas fragments of unstable designs were more geometrically distant from the fragments of natural proteins matching their local sequence ($P < 2 \times 10^{-26}$, fig. S7B). Other metrics were only weakly correlated with success despite substantial variability among designs, including different measures of amino acid packing density and the total Rosetta energy itself. Although local sequence structure agreement and especially buried NPSA are well known to be important for protein stability (1, 9), it is very challenging to determine the precise strength of these contributions at a global level in the complex balance of all the energetic contributions influencing protein structure. Our results directly demonstrate how specific imbalances (underweighting buried NPSA and local sequence structure agreement in the Rosetta energy model and the design procedure) led to hundreds of design failures, and our data and approach provide a new route to refining this balance in biophysical modeling.

**Iterative, data-driven protein design**

We sought to use these findings to increase the success rate of protein design by (i) changing the design procedure to increase buried NPSA, and (ii) reweighting the metrics used to select designs for testing (26). Using the improved design and ranking procedure, we built a second generation of 4150 designs, along with two control sequences per design: a pattern-preserving scrambled sequence as before (now also preserving Gly and Pro positions), and a second control identical to the designed sequence, but with the most buried side chain (according to the design model) replaced with Asp. As in round 1, almost no scrambled sequences had stability scores greater than 1 (our cutoff defining success) despite the increased hydrophobicity of the scrambled sequences (Fig. 2A, round 2). However, a much larger proportion of second-generation designs proved stable: Success for αααα designs improved from 23% to 69%, β/β designs improved from 1% to 11% successful, and we also obtained seven stable αββ designs and one stable βββ design (Fig. 2H). These increases demonstrate how iterative, high-throughput protein design can make concrete improvements in design and modeling. Nearly all stable designs were destabilized via the single buried Asp substitution: The median drop in stability score for these designs was 1.1, and only 33 buried Asp controls had stability scores greater than 1.0, compared with 271 designs (Fig. 2A, round 2). This substantial destabilization from a single designed substitution provides further large-scale evidence that the stable designs fold into their designed structures. We purified and characterized seven second-generation proteins by SEC and CD, all of which (including three αββ designs and one βββ design) were monomeric, displaying their designed secondary structure in CD, and folded cooperatively and reversibly after thermal denaturation (fig. S6B and table S1). Although the αββ and βββ designs were only marginally stable, the second-generation βββ design EHEE_rd1_0005 is, to our knowledge, the most thermostable minimal protein ever found (lacking disulfides or metal coordination): Its CD spectrum is essentially unchanged at 95°C, and its denaturation midpoint concentration (Cm) is above 5 M guanidine hydrochloride (GuHCl) (fig. S6B).

The amount of buried NPSA was the strongest observed determinant of folding stability for second-generation βββ designs (Fig. 2E) and continued to show correlation with stability for second-generation αααα designs (Fig. 2D). The success rate for αααα designs improved in round...
2 at all levels of buried NPSA (compare Fig. 2D with Fig. 2B), indicating that improvement of design properties unrelated to buried NPSA (mainly local sequence structure compatibility) contributed to the increase in success rate along with the increase in NPSA. This also illustrates the coupling between different contributions to stability. Although analyzing single terms makes it possible to identify key problems with the design procedure and imbalances in the energy model, the specific success rates shown in Fig. 2 depend on the overall protein context and are not, on their own, fully general.

To improve the stability of the other two topologies, we built a third generation of designs with even greater buried NPSA, at the cost of increased exposure of hydrophobic surface. This might decrease the solubility of the designs, increased exposure of hydrophobic surface. This might decrease the solubility of the designs, replaced by larger buried NPSA, at the cost of even greater buried NPSA, at the cost of increased exposure of hydrophobic surface. This might decrease the solubility of the designs, replacing the large exposed hydrophobic surface with even greater buried NPSA, at the cost of even greater exposed hydrophobic surface. This might decrease the solubility of the designs, replacing the large exposed hydrophobic surface with even greater exposed hydrophobic surface. This might decrease the solubility of the designs, replacing the large exposed hydrophobic surface with even greater exposed hydrophobic surface. This might decrease the solubility of the designs.
YAP-family WW domains (Fig. 4D and fig. S12) (47, 48). These examples illustrate how our approach enables high-throughput identification of functional residues, even without a functional assay or a protein structure (as in computational approaches (49)), via comparison between stability data and residue conservation.

**Stability measurement of all known small protein domains**

How stable are these designed proteins relative to naturally occurring proteins? To examine this, we synthesized DNA encoding (i) all 472 sequences in the Protein Data Bank (PDB) between 20 and 50 residues in length and containing only the 19 non-Cys amino acids, and (ii) one representative for all 706 domains meeting these criteria in the Pfam protein family database. These DNA sequences were prepared by reverse translation in an identical manner to the designs (26). We included this DNA (and DNA for all stable designs from rounds 1 to 3) in the library containing our fourth-generation designs to facilitate a head-to-head comparison. The large majority of these natural proteins successfully displayed on yeast (92% each for PDB and Pfam sequences), which was comparable to the fourth-generation buried Asp mutants (also 92%) but lower than fourth-generation scrambled sequences (96%) and fourth-generation designs (99%). The most resistant overall sequence (measured by stability score) was a C-terminal coiled-coil domain from a TRP channel (3HRO, stability score 1.03). This protein is likely stabilized by intersubunit interactions made possible by assembly on the yeast surface (50). Of the 100 unique, monomeric sequences with PDB structures, the most protease-resistant was a peripheral subunit binding domain (alpha topology) from the thermophile *Bacillus stearothermophilus* (2PDD, stability score 1.48), which has been studied as an ultrafast-folding protein (4, 5). A total of 774 designed proteins had higher stability scores than this most protease-resistant natural monomeric protein. As illustrated in Fig. 5, the number of stable proteins we discovered exceeds the number of natural proteins in the PDB (monomeric or not) in this size range by a factor of 50.

![Figure 4](http://science.sciencemag.org/content/357/6356/168/F4.large.jpg)

**Fig. 4. Comprehensive mutational analysis of stability in designed and natural proteins.** (A) Average change in stability due to mutating each position in 13 designed proteins, depicted on the design model structures. Positions where mutations are most destabilizing are colored yellow and shown in stick representation; positions where mutations have little effect are colored blue. Each protein's color scale is different to emphasize the relative importance of positions; see fig. S10 for full data for all proteins. (B to D) As in (A) for native proteins, with conserved residues not contributing to stability colored red. (B) Villin HP35. In red, Trp64, Lys70, Lys76, and Phe76 (HP35 consists of residues 42 to 76) have little effect on stability but are conserved for function (F-actin binding). (C) Pin1 WW domain, shown bound to a doubly phosphorylated peptide. In red, Ser16 is conserved and critical for function but is destabilizing relative to mutations at that position. (D) hYAP65 L30K, shown bound to a Smad7-derived peptide. In red, His32, Thr37, and Trp39 form the peptide recognition motif and are conserved but unimportant for stability. (E to L) Average stability effect of each amino acid at different categories of surface positions, in units of stability score (positive = stabilizing, negative = destabilizing). The average stability of all amino acids in each panel was set to zero. The number of individual positions examined in each category is listed in parentheses with the category name. The average stability effect of the original “wild-type” designed residue (unique to each particular site within a category) is shown by a black star. Error bars indicate the 50% confidence interval for the average stability effect, calculated using bootstrapping. See (26) for a full description of the analysis.
structures. We have entered a new era of iter-
minants of protein stability in more complex
DNA synthesis technology continues to improve,
translocation and endosomal escape (small size may also help to promote membrane
meric proteins found in the PDB, making them
nation, should have numerous applications in
teins, stable without disulfides or metal coordi-
plausible but unstable protein structures like our
PDB structures (lacking disulfides) in each bin is shown by a black bar.

Fig. 5. Comparison of naturally occurring and designed protein stability. Designed and naturally occurring proteins are separated into bins by stability score (y axis). The total number of designed proteins in each bin is shown by the colored bar, subdivided by topology from left to right as follows: αααα (green), ββββ (blue), αβαβ (violet), βαβα (red). The total number of naturally occurring proteins with PDB structures (lacking disulfides) in each bin is shown by a black bar.

**References and Notes**

26. See supplementary materials.

**Acknowledgments**

Supported by the Howard Hughes Medical Institute (D.B.) and the Natural Sciences and Engineering Research Council of Canada (C.H.A.). G.J.R. is a Merck Fellow of the Life Sciences Research Foundation. C.H.A. holds a Canada Research Chair in Structural Genomics. We thank S. Rettie for mass spectrometry support; C. Lee for deep sequencing support; S. Ochinokr for assistance quantifying sequence conservation; V. Nguyen, A. Yehdego, T. Howard, and K. Lau for assistance with protein purification; and H. Gelman and many other members of the Baker lab for helpful discussions. This work was facilitated by the Hyak supercomputer at the University of Washington and by donations of computing time from the University of Washington that covers (i) the method described in this work for computational designing and experimentally verifying stable mini-proteins, and (ii) the 4000 most stable protein structures designed in the work. Author contributions: G.J.R. designed the research, the experimental approach, and the proteins; G.J.R., T.M.C., I.G., S.H., L.C., R.R., and A.C. performed experiments; all authors analyzed data; G.J.R., A.F., and V.K.M. contributed new computational tools; C.H.A. and D.B. supervised research; and G.J.R. and D.B. wrote the manuscript.

**Supplementary Materials**

www.sciencemag.org/content/357/6347/168/suppl/DC1

Materials and Methods Supplementary Text
Fig. S1 to S3 Tables S1 to S3 References (54–88)

28 February 2017; accepted 9 June 2017.
Global analysis of protein folding using massively parallel design, synthesis, and testing

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith and David Baker

Science 357 (6347), 168-175.
DOI: 10.1126/science.aan0693

Exploring structure space to understand stability

Understanding the determinants of protein stability is challenging because native proteins have conformations that are optimized for function. Proteins designed without functional bias could give insight into how structure determines stability, but this requires a large sample size. Rocklin et al. report a high-throughput protein design and characterization method that allows them to measure thousands of miniproteins (see the Perspective by Woolfson et al.). Iterative rounds of design and characterization increased the design success rate from 6 to 47%, which provides insight into the balance of forces that determine protein stability.

Science, this issue p. 168; see also p. 133

ARTICLE TOOLS
http://science.sciencemag.org/content/357/6347/168

SUPPLEMENTARY MATERIALS
http://science.sciencemag.org/content/suppl/2017/07/12/357.6347.168.DC1

RELATED CONTENT
http://science.sciencemag.org/content/sci/357/6347/133.full

REFERENCES
This article cites 86 articles, 15 of which you can access for free
http://science.sciencemag.org/content/357/6347/168#BIBL

PERMISSIONS
http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service