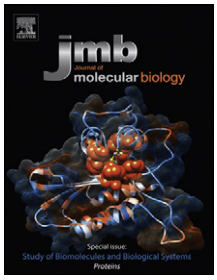

De Novo Proteins with Life-Sustaining Functions Are Structurally Dynamic

Grant S. Murphy, Jack B. Greisman and Michael H. Hecht

Department of Chemistry, *Princeton University, Princeton, NJ 08540, USA*

Correspondence to Michael H. Hecht: *hecht@princeton.edu*
http://dx.doi.org/10.1016/j.jmb.2015.12.008
Edited by A. Skerra

Abstract

Designing and producing novel proteins that fold into stable structures and provide essential biological functions are key goals in synthetic biology. In initial steps toward achieving these goals, we constructed a combinatorial library of *de novo* proteins designed to fold into 4-helix bundles. As described previously, screening this library for sequences that function *in vivo* to rescue conditionally lethal mutants of *Escherichia coli* (auxotrophs) yielded several *de novo* sequences, termed SynRescue proteins, which rescued four different *E. coli* auxotrophs. In an effort to understand the structural requirements necessary for auxotroph rescue, we investigated the biophysical properties of the SynRescue proteins, using both computational and experimental approaches. Results from circular dichroism, size-exclusion chromatography, and NMR demonstrate that the SynRescue proteins are α-helical and relatively stable. Surprisingly, however, they do not form well-ordered structures. Instead, they form dynamic structures that fluctuate between monomeric and dimeric states. These findings show that a well-ordered structure is not a prerequisite for life-sustaining functions, and suggests that dynamic structures may have been important in the early evolution of protein function.

Introduction

The two central challenges of protein design are (i) to devise novel amino acid sequences that fold into stable three-dimensional structures and (ii) to devise sequences that perform chemically and/or biologically significant functions. Early work in protein design began approximately 25 years ago, with attempts to design 4-helix bundles [1,2]. Those pioneering studies focused exclusively on folding and stability, and they paid little attention to protein function. This seemed reasonable at the time because it was assumed that achieving a well-ordered structure was an essential prerequisite for protein function. Because of this assumption, it was only in recent years, as the design of stably folded structures achieved some level of success [3–8], that protein designers began to consider the possibility of devising novel proteins that bind targets and/or catalyze reactions [9–12].

The presumption that uniquely folded structures are essential for function arose from the pioneering achievements of structural biology. The first crystal structures, solved more than half a century ago, revealed ordered structures with well-defined active sites that accounted for their biochemical functions [13]. After observing such structures, it is not surprising that researchers assumed that a well-ordered structure was a prerequisite for a well-defined function. Indeed, these early findings led to a central paradigm of structural biology: amino acid sequence determines three-dimensional structure, and structure—typically denoting a well-ordered structure—determines function.

In recent years, however, numerous studies have demonstrated that many natural proteins responsible for essential cellular functions are, in fact, intrinsically disordered and/or dynamic [14,15]. In light of these findings, it may be time to reconsider assumptions about the relationship between well-ordered structures and biological function—both for naturally evolved proteins and for proteins designed *de novo*.

In the current study, we question these assumptions by probing the structural and biophysical properties of several α-helical proteins, which were designed *de novo* in our laboratory and shown previously to function *in vivo* by providing life-sustaining activities

in *Escherichia coli* [16]. Using a range of experimental techniques, we probe whether these functional *de novo* proteins fold into well-ordered, kinetically stable structures or, alternatively, fluctuate between dynamic states.

The *de novo* α-helical proteins that are the subject of the current study were drawn from a large combinatorial library of binary patterned sequences that we described previously [16–18]. Briefly, binary patterning is a strategy for protein design, which is built on the premise that the overall structure of a protein can be specified by designing the sequence periodicity of polar and nonpolar amino acids to match the structural periodicity of the desired secondary structure. Thus, a pattern that places a nonpolar amino acid every 3 or 4 residues along a sequence would match the structural repeat of 3.6 residues per turn of a canonical α-helix and thereby would generate an amphiphilic α-helical segment. When four such helices are linked together, the hydrophobic effect drives them to pack against one another, thereby forming a 4-helix bundle with nonpolar residues pointing toward the protein core and polar residues exposed to solvent (Fig. 1a). Since only the *type* of residue—polar *versus* nonpolar—is designed explicitly, the strategy is inherently binary. However, because the *identities* of the polar and nonpolar side chains are *not* specified, the strategy is inherently combinatorial and facilitates the construction of vast libraries of novel sequences.

The combinatorial diversity of the protein library is encoded at the DNA level by using degenerate codons, such as NTN (N = A, T, C, or G) to encode five nonpolar amino acids (Phe, Leu, Ile, Met, and Val) and VAN (V = A, C, or G) to encode six polar amino acids (His, Glu, Gln, Asp, Asn, and Lys). These degenerate codons can be assembled in a pattern compatible with the desired structure to produce a collection of synthetic genes, which can be translated in *E. coli* to produce a large library of *de novo* proteins.

Previously, we reported the construction of three binary patterned libraries of sequences designed to fold into 4-helix bundles [17,19,20]. The sequences in these libraries do not share homology with naturally occurring proteins. They were not selected by eons of evolution, and they may share features with primordial sequences that existed in the early history of life on earth.

Previous studies of proteins from these binary patterned libraries showed that many of the sequences fold into stable structures [20]. Three structures were determined by NMR or crystallography to reveal 4-helix bundles with hydrophobic interiors and polar surfaces, as envisioned by the binary patterned design. Two proteins from our second-generation library formed monomeric 4-helix bundles [4,21], while an X-ray structure solved from a sequence from the third-generation library revealed a domain-swapped dimer [22]. We have also identified *de novo* proteins from these libraries that bind small molecules, including drugs and cofactors [18,23]. Furthermore, we identified sequences that possess weak catalytic activity for simple reactions and substrates, such as the hydrolysis of *p*-nitrophenyl esters [18].

The results summarized in the previous paragraph demonstrated that proteins from binary patterned libraries possess structural and functional properties *in vitro* resembling those of natural proteins. More recently, we have become interested in the possibility of designing collections of novel sequences as an initial step toward constructing artificial “proteomes”. This interest led to experiments probing the ability of our novel sequences to provide essential functions *in vivo*. Since the proteins in our libraries were designed for structure, but not explicitly designed for any particular function, we used unbiased high-throughput genetic selections to search for novel sequences that functioned *in vivo*. These selections relied on a series of *E. coli* auxotrophs: strains that are deleted for individual genes that encode enzymes necessary for survival on minimal medium. In a typical auxotroph rescue experiment, an *E. coli* auxotrophic strain was transformed with a binary patterned library encoding 10^6 *de novo* proteins. In most cases, the auxotroph was not rescued by sequences from our library; however, four auxotrophic strains of *E. coli* were rescued by sequences from our third-generation binary patterned library [16]. The four rescued auxotrophic strains are deleted for a range of functions: Δ*fes* is missing enterobactin esterase, Δ*ilvA* is missing threonine deaminase, Δ*serB* is missing phosphoserine phosphatase, and Δ*gltA* is missing citrate synthase. In all, more than 20 *de novo* sequences were found to rescue one of these four deletion strains. We denote these novel sequences the SynRescue proteins because they are *syn*thetic (not derived from nature) and they *rescue* the given deletion strain. Individual proteins are named SynΔ-strain#, such that SynFes2 is the second *de novo* protein identified that rescued Δ*fes*.

It is tempting to assume that the SynRescue proteins rescue the deletion strains in a direct manner by performing the same biochemical activity as the deleted protein. However, this need not be the case. It is also possible for a SynRescue protein to compensate for a deleted protein by increasing the expression, enhancing the activity, or altering the specificity of an endogenous *E. coli* protein. Irrespective of the mechanism of rescue, structural and biophysical characterization of the SynRescue proteins may help elucidate their functions.

The SynRescue proteins also present an unusual opportunity to revisit the relationship between well-ordered structure and biological function. Moreover, because these sequences were devised *de novo* in the laboratory, we can ask whether uniquely folded three-dimensional structures are essential for function *in vivo* in a system that is not

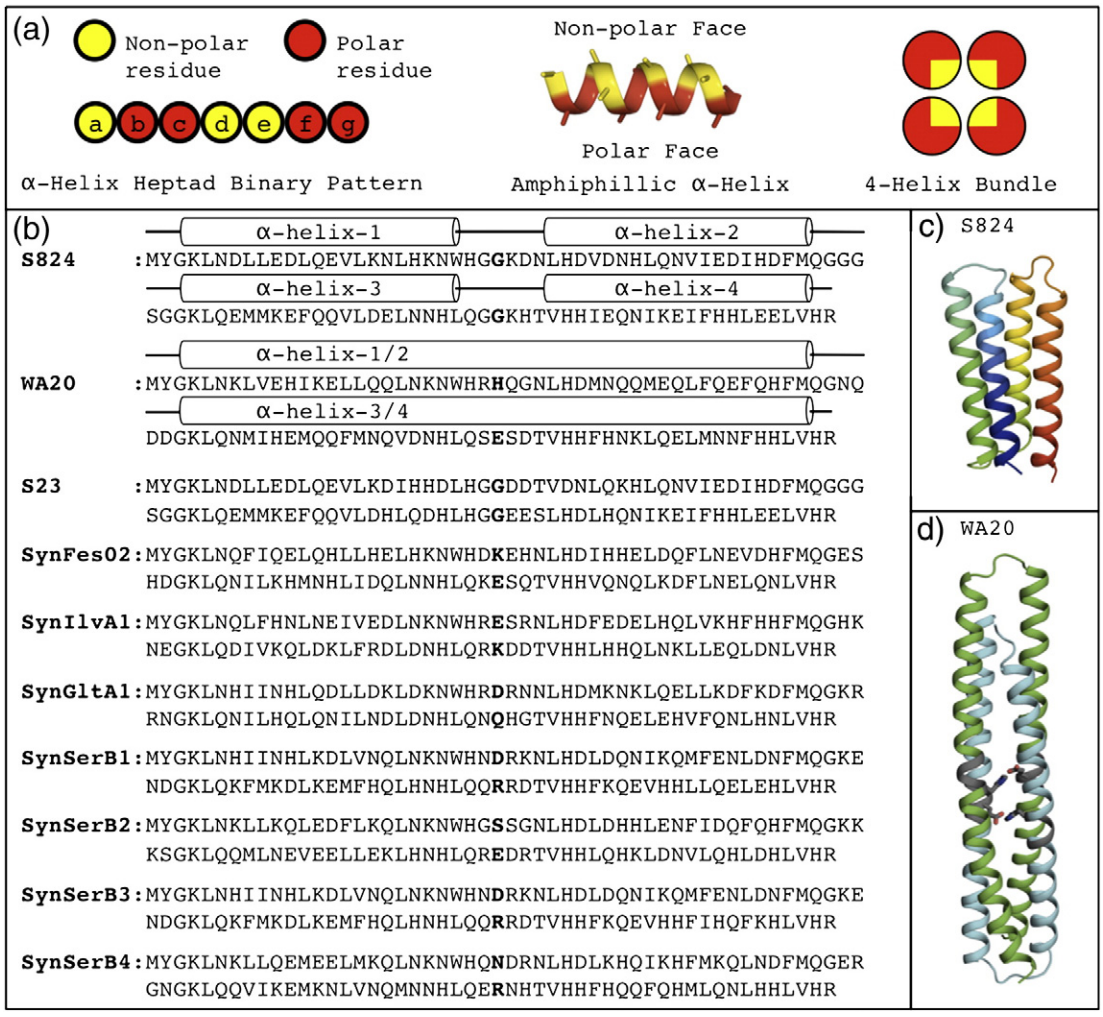

Fig. 1. The binary code strategy for protein design and the sequences of the characterized proteins. (a) The binary code strategy designs amino acid sequences by placing polar (red) and nonpolar (yellow) residues to match the structural periodicity of an α-helix. Thus, helix heptad positions a, d, and e are designed to be nonpolar, while positions b, c, f, and g are polar. This binary patterning can direct four amphiphilic α-helices to assemble into a 4-helix bundle. (b) The sequences of the control proteins of S824 and WA20 are shown with their α-helices shown as cylinders. (c) Structure of S824 [4]. (d) Structure of WA20 [22]. Protein S824 forms a monomer and WA20 forms an extended domain-swapped dimer. In WA20, the buried polar amino acids H26 and E78, which form a set of buried hydrogen bonds, are shown as sticks and the positions 26 and 78 are boldfaced for all sequences.

biased by eons of evolutionary history. To address these questions, we investigated the biophysical properties of the SynRescue proteins, using both computational and experimental approaches. Results from circular dichroism (CD), size-exclusion chromatography (SEC), and NMR demonstrate that the SynRescue proteins are α-helical and relatively stable. Surprisingly, however, they do not form well-ordered structures. Instead, they form dynamic structures that fluctuate between monomeric and dimeric states. These findings show that well-ordered structure is *not* a prerequisite for function *in vivo*, and they suggest that dynamic structures may have been important in the early evolution of protein function.

Results

The SynRescue proteins

For this investigation, we explored the biophysical and structural properties of seven SynRescue proteins: SynFes2, which rescues Δ*fes*; SynGltA1, which rescues Δ*gltA*; SynIlvA1, which rescues Δ*ilvA*; and SynSerB1, SynSerB2, SynSerB3, and SynSerB4, which rescue Δ*serB*. We compared their properties to three control proteins S824, S23, and WA20. The proteins S23 and S824 are sequences from the second-generation library (hence the “S”

prefix). We previously reported the solution NMR structure of S824, which confirmed that it folds into a 4-helix bundle, as designed previously [4]. S23 was shown previously to be a monomeric molten globule α-helical protein [20].

S824 was the template sequence for the binary pattern and constant regions of the third-generation library [17]. The SynRescue sequences are all members of the third-generation library, and they have between 42% and 51% sequence identity with S824. The protein WA20 is also a member of the third-generation library. We recently solved the crystal structure of WA20 to 2.2 Å, which revealed a 4-helix bundle comprising a domain-swapped dimer [22]. Figure 1 shows the sequences of the SynRescue proteins; the control proteins S23, S824, and WA20 (1B); and the experimentally determined structures of S824 (1C) and WA20 (1D).

Computational structure prediction

We performed computational structure prediction simulations for each of the SynRescue proteins and the control proteins S23, S824, and WA20, using the macromolecular modeling software Rosetta, which has been shown to accurately predict the structures of many small proteins (<150 residues) [24]. The NMR solution structure of S824 has previously been solved (PDB code 1p68), and S824 is an extremely stable and well-ordered monomeric 4-helix bundle [4]. We attempted to computationally predict the structure of S824 as a positive control for Rosetta's ability to predict the structure of *de novo* sequences, not designed in Rosetta and, which have amino acid distributions that differ significantly from natural proteins (e.g., these sequences do not contain alanine or proline). Supplemental Figure 1 shows a plot of the root-mean-square deviation (RMSD) *versus* total Rosetta energy for the S824 structure prediction. In an ideal case, a single "folding funnel" would be observed at low RMSD and low Rosetta energy [24]; however, the plot for S824 shows several funnels with approximately equal energies. While RMSD space is highly multidimensional, the lowest energy models in each funnel correspond to the different possible topologies of a 4-helix bundle. Although the Rosetta simulation samples the experimentally determined topology, the energy function is not able to accurately identify the correct structure of S824.

For each of the four "folding funnels", we used the experimentally determined nuclear Overhauser effect (NOE) distance constraints from protein S824 to calculate the number of violations for each model structure. Only models with the same topology as the S824 NMR solution structure, left-handed 4-helix bundles (green funnel in Supplemental Fig. 1), satisfied the NOE distance constraints. Models from the other three topologies have hundreds of long-range NOE distance violations, confirming that the only structure compatible with these chemical shifts and NOE constraints is the experimentally determined structure.

We performed similar simulations for the SynRescue proteins, and similar to S824, they showed multiple funnels with similar energies. Investigation of the lowest energy models did not indicate which fold, if any, would be the true structure (Supplemental Fig. 2 shows the prediction results for the SynRescue sequences). We also performed structure prediction simulations for WA20. Since the X-ray crystal structure of WA20 is a homodimer, we used Rosetta's fold-and-dock protocol [25]. The fold-and-dock structure prediction results for WA20 also showed multiple folding funnels with approximately equal energies. The lowest energy models in each funnel correspond to different arrangements of a helix–turn–helix homodimer. Again, the simulation sampled the experimentally determined topology; however, the Rosetta energy function did not identify models with the topology of the X-ray crystal structure as the lowest energy models (Supplemental Fig. 3).

These simulations demonstrate that, for S824 and WA20, Rosetta's monomer and oligomer structure prediction methods sample the correct conformational space but the energy function does not identify the experimentally determined structure as having the lowest energy. This could occur for several reasons: (1) the sequences predicted here have features that are not common in natural proteins or in Rosetta *de novo* designed proteins, such as they do not contain the amino acids alanine, proline, and cysteine, and they have unusual amino acid distributions (e.g., overrepresentation of histidine). Since many terms in the Rosetta energy function are trained on high-resolution X-ray crystal structures of natural proteins and the Rosetta reference energy is trained specifically to recapitulate "natural" amino acid distributions, the Rosetta energy function may not accurately represent the energies of these binary patterned proteins. (2) The actual physical energy differences between the structures sampled in the Rosetta simulations may be small and within the error of the Rosetta energy function. (3) In the cases of WA20 and the SynRescue proteins, we have not solved their NMR solution structures; thus, the Rosetta simulations may be correct in suggesting these sequences sample multiple topologies.

Protein expression and purification

We expressed and purified the control proteins S23, S824, and WA20 and the seven SynRescue proteins. The control proteins S23, S824, and WA20 express and purify with high yield. However, some SynRescue sequences express and purify much more readily than others (see the methods section for details). In all cases, it was possible to

generate pure protein (>95% by SDS-PAGE) at concentrations of at least 200 μM for biophysical and structural characterization.

The SynRescue proteins form α-helical secondary structure

CD measurements of the SynRescue and control proteins revealed canonical spectra with minima at 208 and 222 nm, thereby demonstrating that, as expected from their binary patterned design, the proteins are predominantly α-helical (Fig. 2a).

Most of the SynRescue proteins display similar levels of α-helical content, except for SynGltA1, which shows ~50% of the α-helical content of the other proteins. For helical proteins, the ratio of ellipticity at 222 nm relative to 208 nm indicates the amount of supercoiling. A 222/208 ratio greater than 1.0 is consistent with coiled-coil structures, whereas values between 0.9 and 1.0 indicate assemblies of nonsupercoiled helices, and values less than 0.9 suggest independent helices [26]. The control protein WA20 has a 222/208 ratio of 1.2 indicating that it is supercoiled in solution, as expected from its crystal structure, where the domain-swapped dimeric bundle is twisted by ~90° along its long axis. The protein S824 has a 222/208 ratio of 0.98 indicating that it is not extensively supercoiled, consistent with the NMR structure. The SynRescue proteins also have 222/208 ratios of ~1, indicating that they are not highly supercoiled (Supplemental Table 1).

Thermal stability

To assess the thermal stability of the SynRescue proteins, we monitored ellipticity at 222 nm as a function of temperature. The control proteins S824 (Fig. 2b, green circle) and WA20 (Fig. 2b, black pentagon) are thermostable, with unfolding midpoints of >100 °C and 80 °C, respectively. All of the SynRescue proteins are also stable, with denaturation midpoints between 50 and 90 °C (Supplemental Table 1). The thermal denaturations for SynSerB1 (yellow-filled squares) and SynGltA1 (orange-filled triangles) are shown in Fig. 2b and are representative of the extremes of the SynRescue proteins. The denaturation curves of the SynRescue proteins have a range of cooperativities, with some being barely cooperative (SynFes2) and others being modestly cooperative (SynIlvA1).

Thermal denaturations of the SynRescue and control proteins were thermodynamically reversible: after cooling to the original temperature, followed by a period of equilibration, ellipticity at 222 nm regained 95–100% of the original native values. Although all of the samples display thermodynamic reversibility, the kinetics of refolding differed among the various sequences. Protein S824, which is known to form a well-ordered monomeric 4-helix bundle [4], refolded relatively rapidly with its renaturation curve nearly superimposable on its denaturation curve. In contrast, WA20, which is known from crystallography to form a domain-swapped

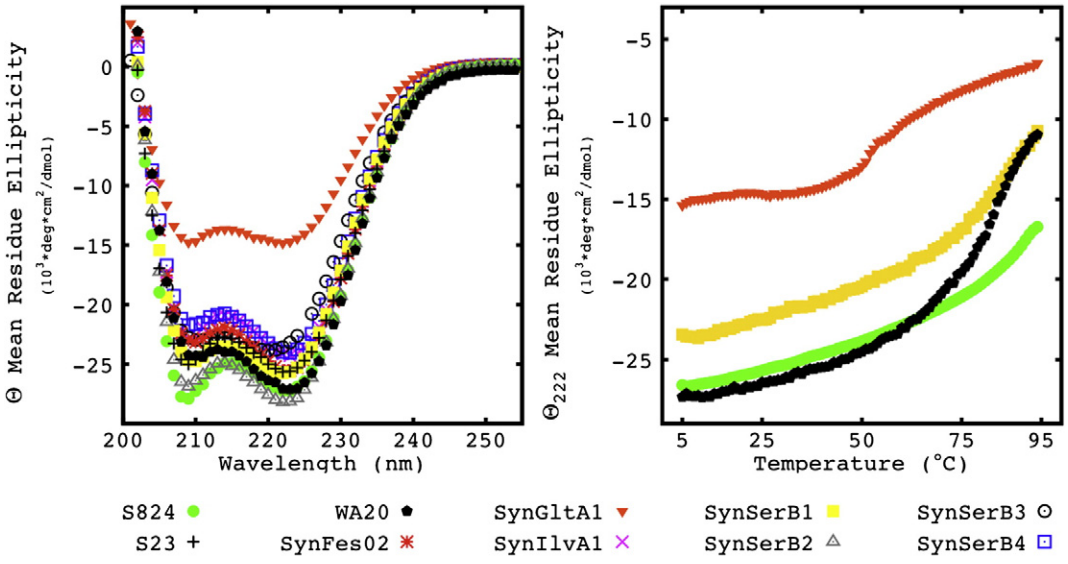

Fig. 2. The SynRescue proteins are helical and stable. (a) Far-UV CD spectra. The control proteins S824 (green circle), S23 (plus sign), and WA20 (black pentagon) and the rescue proteins SynFes02 (red star), SynGltA1 (orange down triangle), SynIlvA1 (purple X), SynSerB1 (yellow-filled square), SynSerB2 (gray up triangle), SynSerB3 (black open circle), and SynSerB4 (blue open square) display CD spectra consistent with α-helical structures, with prominent minima at 208 nm and 222 nm. (b) Thermal denaturation. The SynRescue proteins display a range of thermal stabilities. SynGltA1 (orange down triangle) has the lowest midpoint and SynSerB1 (yellow-filled square) has one of the highest midpoints. The control protein S824 (green-filled circle) is shown for comparison as an extremely stable monomer. The dimer control protein WA20 (black pentagon) is shown and behaves similar to SynSerB1.

dimer [22], refolded more slowly, with its renaturation lagging behind the original denaturation curve. The SynRescue proteins displayed delayed renaturation, similar to that observed for the WA20 dimer (Supplemental Fig. 4).

NMR spectroscopy

To probe the structural properties of the SynRescue proteins, we recorded their $^{1}H^{15}N$ heteronuclear single quantum coherence (HSQC) NMR spectra (Fig. 3). In such spectra, a monodisperse, well-folded protein is expected to show a cross-peak for each backbone NH and a pair of cross-peaks for each asparagine and glutamine side chain. The control monomeric protein, S824, yields such a spectrum, with abundant and well-resolved peaks (Fig. 3a). In contrast, a molten globule protein, which is compact but dynamic, would be expected to produce a spectrum with limited chemical shift dispersion. The control monomeric protein, S23, yields a spectrum consistent with the molten globule state (Fig. 3b). The control protein WA20, which forms a dimer in solution and in its X-ray crystal structure, has spectra consistent with a molecule undergoing exchange between multiple states on the timescale of the NMR experiment. The peaks of WA20's $^{1}H^{15}N$ HSQC are broad, low intensity, and poorly resolved.

The $^{1}H^{15}N$ HSQC NMR spectra of the SynRescue proteins resemble *neither* the well-folded *nor* molten globule monomeric control proteins but instead resemble the spectra of WA20. The spectra of the SynRescue proteins show peaks with low intensity and broad linewidths. In some cases, the SynRescue spectra have numerous broad, low-intensity backbone NH peaks (e.g., SynIlvA1 and SynSerB4 in Fig. 3d and e), while in other cases the spectra display relatively few broad, low-intensity peaks (SynGltA1 in Fig. 3f and SynSerB1, SynSerB2, SynSerB3, and SynFes2 in Supplemental Fig. 5). These spectra are consistent with dynamic structures undergoing exchange on a range of intermediate timescales. Given the nature of the $^{1}H^{15}N$ HSQC spectra and the slow reversibility of refolding observed in the thermal denaturations, we considered the possibility that the SynRescue proteins might be undergoing exchange between monomeric and oligomeric states on an intermediate timescale.

In some cases, it is has been possible to assign or partially assign the chemical shifts of proteins undergoing exchange on an intermediate timescale. However, considering the quality of the SynRescue $^{1}H^{15}N$ HSQC spectra and similar data quality in other experiments traditionally used in backbone and side-chain assignment ($^{1}H^{13}C$ HSQC, HNCA, HNCO, HNCACB, and HNCACO were collected for several SynRescue proteins; data not shown), we concluded that it would not be possible to assign or even partially assign the backbone or side-chain chemical shifts of the SynRescue proteins using traditional methods and the conditions tested. While we could not determine the structures of the SynRescue proteins by solution NMR, we still wanted to investigate the oligomeric state of the SynRescue proteins. Therefore, we probed the solution state of these proteins by SEC.

Size-exclusion chromatography

The oligomeric states of the SynRescue proteins were assessed by SEC. Because SEC is a nonequilibrium method, the apparent molecular weight of a protein undergoing monomer/oligomer exchange on an intermediate timescale will be influenced by the time spent on the column and by the flow rate and size of the column. Therefore, we measured the apparent molecular weights of the *de novo* proteins using columns of three different sizes: (i) an S75 5/150 analytical column with a 3-mL bed volume, (ii) an S75 10/300 semipreparative column with a 24-mL bed volume, and (iii) an S75 26/600 preparative column with a 318-mL bed volume.

The well-folded monomer S824, the molten globule monomer S23, and the domain-swapped dimer WA20 provide appropriate controls for this experiment. The molecular masses of S824 and S23, calculated from their amino acid sequences, are both 11.9 kDa. In SEC experiments, both proteins run at ~12 kDa on all three columns, confirming that these proteins exist in solution as monodisperse monomeric globular structures (Fig. 4a and Table 1).

The other control protein, WA20, has a covalent molecular mass of 12.5 kDa, calculated from its amino acid sequence. The crystal structure of WA20 shows a domain-swapped dimer, and the expected molecular mass of this dimer would be 25 kDa. However, the dimer seen in the crystal structure is elongated. This is because turn 1 and turn 3 of the intended design did not form and instead continue helix 1 into helix 2, as well as helix 3 into helix 4 (Fig. 1a and c). This causes WA20 to be shaped more similar to a rod than a sphere. Since SEC separates proteins based on their hydrodynamic radii [27], the rod-shaped structure of WA20 would be expected to run through SEC columns with an apparent molecular mass that is larger than would be observed for a more spherical, 25-kDa protein.

On the smallest SEC column (5/150), WA20 runs with an apparent molecular mass of 31.4 kDa, which is ~2.5 times its expected monomer molecular weight and is consistent with its elongated structure and larger hydrodynamic radius. However, on the medium-sized column (10/300), the apparent molecular mass of WA20 is shifted to 25.2 kDa, which is ~2 times its covalent molecular weight. Finally, with the use of the largest column (26/600), the apparent molecular mass of WA20 is further shifted to 20.6 kDa, which is only 1.7 times WA20's covalent

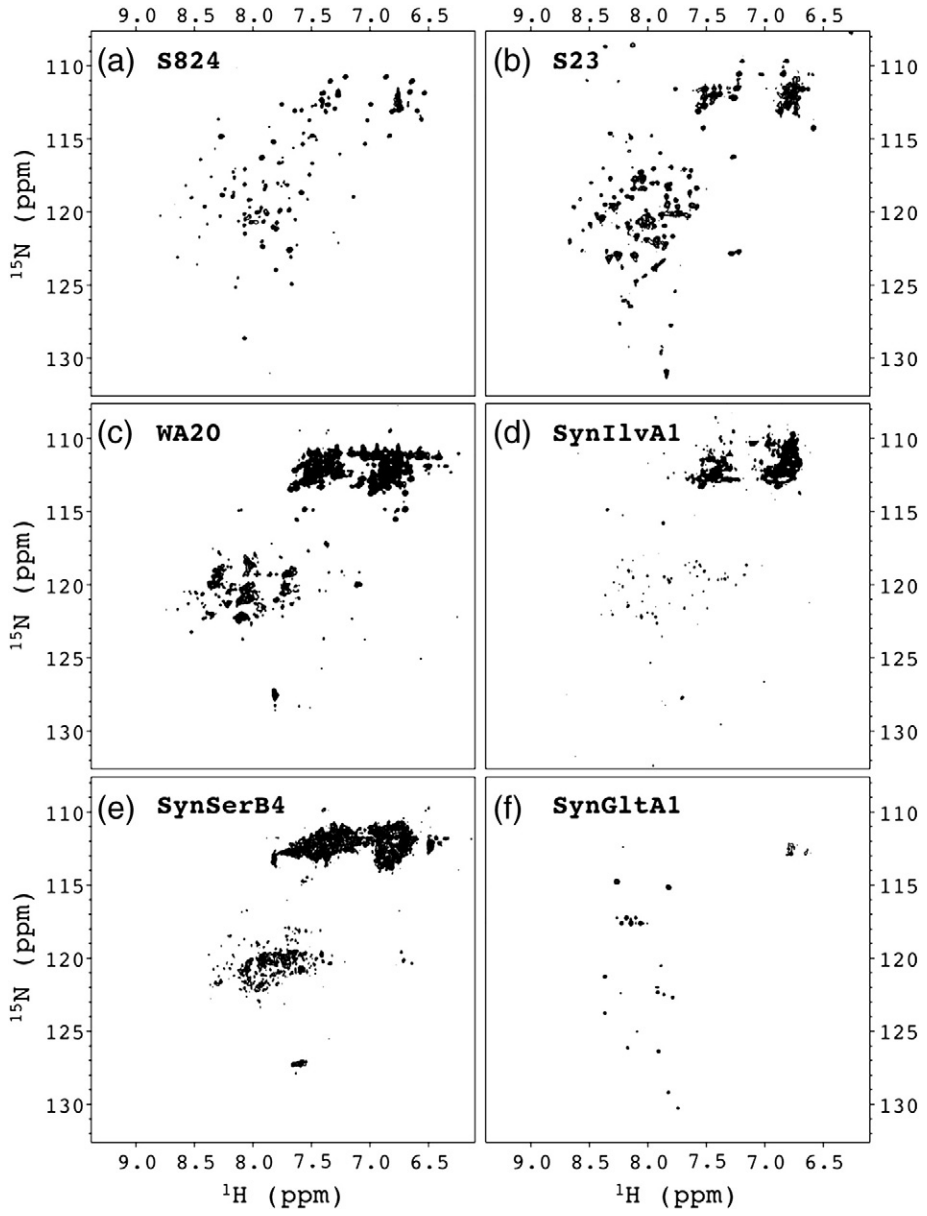

Fig. 3. $^{1}H^{15}N$ HSQC NMR spectra indicate that the SynRescue proteins are dynamic. (a) The spectrum of the well-folded *de novo* protein S824 shows intense peaks with unique chemical shifts for each backbone NH and Asn and Gln side-chain NH. (b) The spectrum of the control molten globule S23 shows numerous peaks but with many overlapping chemical shifts. (c) The spectrum of the extended dimer WA20 shows numerous broad, low-intensity peaks consistent with a structure undergoing intermediate exchange. (d and e) The spectra of SynIlvA1 and SynSerB4 show numerous broad, low-intensity peaks indicating that they are dynamic. (f) The spectrum of SynGltA1 shows approximately one-fifth of the expected backbone peaks, indicating that it is primarily unfolded or extremely dynamic.

monomer weight. These results suggest that, during the longer runs on the larger SEC column, WA20 dissociates from its dimeric structure.

Figure 4 compares the apparent molecular weights—on all three SEC columns—of the control proteins S824 and WA20, with representative SynRescue proteins SynFes2 and SynGltA1. (The other SynRescue proteins display behaviors between the extremes of SynFes2 and SynGltA1 and are summarized in Table 1.) The apparent molecular mass of SynFes2 is highly dependent on the column size, running at 27.8 kDa, 23.5 kDa, and 20.6 kDa for the small, medium, and large columns, respectively. On the smallest column, the apparent molecular weight of SynFes2 is 2.3 times its monomer weight. We interpret this as indicating that SynFes2

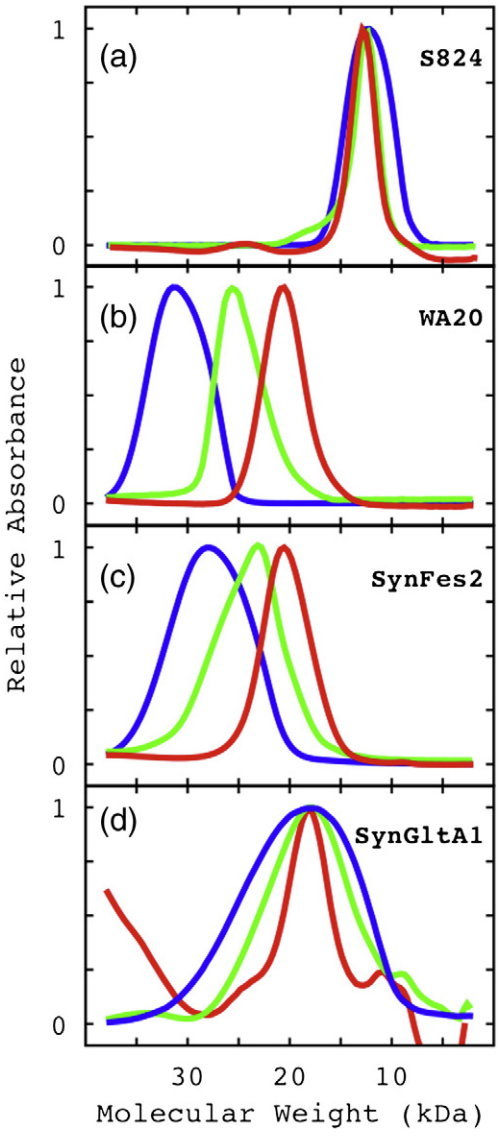

Fig. 4. Apparent molecular masses of the SynRescue proteins. (a) The monomeric control protein, S824, has apparent molecular masses of ~12 kDa on three size-exclusion columns from smallest to largest: S75 5/150 (blue), S75 10/300 (green), and S75 26/600 (red). (b) The dimer control protein, WA20, has apparent molecular masses of 31 kDa (5/150), 25 kDa (10/300), and 20 kDa (26/600). The SynRescue proteins SynFes2 and SynGltA1 are presented as representatives of the extremes of the SynRescue protein's behaviors. SynFes2 has apparent molecular masses of 28 kDa (5/150), 24 kDa (10/300), and 20 kDa (26/600), respectively, and SynGltA1 where the apparent molecular mass on all three columns is ~18 kDa.

forms an extended dimer similar to WA20. This assumption is strengthened by the finding that the apparent molecular weights of SynFes2 on the medium and large columns are similar to those of WA20.

Table 1. Apparent molecular masses of the SynRescue and control proteins.

Construct	MW_{AA} (kDa)	$MW_{5/150}$ (kDa)	$MW_{10/300}$ (kDa)	$MW_{26/600}$ (kDa)
S23	11.9	12.3	12.4	12.5
S824	11.9	12.3	12.4	12.5
WA20	12.5	31.4	25.2	20.6
SynFes02	12.5	27.8	23.5	20.6
SynGltA1	12.5	18.6	18.2	18.4
SynIlvA1	12.6	26.0	23.1	19.5
SynSerB1	12.6	28.8	24.6	19.8
SynSerB2	12.3	27.0	22.0	18.6
SynSerB3	12.7	28.8	21.8	20.1
SynSerB4	12.6	27.5	24.0	17.4

The apparent molecular masses of the SynRescue and control proteins were determined using three size-exclusion columns: an analytical S75 5/150 ($MW_{5/150}$), a semipreparative S75 10/300 ($MW_{10/300}$), and a large preparative S75 26/600 ($MW_{26/600}$). Comparison of the expected monomer molecular mass (MW_{AA}) as calculated from the amino acid sequence with the experimentally determined apparent molecular masses shows that the SynRescue proteins have apparent molecular masses that are consistent with the formation of weakly associated dimers similar to the known dimer WA20.

For SynGltA1, the situation is somewhat different. The apparent molecular mass of SynGltA1 does not depend on column size; it runs at ~18 kDa on all three columns. We interpret this to indicate that either SynGltA1 forms a very weakly associating dimer or it forms an extended monomer. It seems unlikely that SynGltA1 forms a canonical 4-helix bundle (similar to S824 in Fig. 1b) because we do not observe an apparent molecular weight consistent with that structure.

We also evaluated the apparent molecular weight of the control proteins and the SynRescue proteins as a function of protein concentration on the analytical S75 5/150. We tested the proteins at the same concentration used in the NMR, ≥200 μM, and also diluted them to 30 μM. In the concentration range tested, the apparent molecular weight was independent of protein concentration.

The results of the SEC experiments for the remaining SynRescue proteins are summarized in Table 1. All together, we take these results to indicate that the SynRescue proteins form extended helical monomers that assemble into extended dimer structures similar to the crystal structure of WA20. Most importantly, these data, together with NMR spectra, indicate that the SynRescue proteins do not form well-folded or molten globule monomeric structures such as S824 or S23. Instead, the SynRescue proteins appear to fluctuate between monomeric and dimeric α-helical bundles similar to WA20.

Discussion

We investigated the biophysical and structural properties of several *de novo* proteins that were shown previously to provide activities capable of sustaining the growth of living cells. We determined that the SynRescue proteins are α-helical and thermostable and that they denature reversibly. However, $^{1}H^{15}N$ HSQC NMR experiments demonstrate that their structures are dynamic and undergo kinetic exchange on an intermediate timescale. SEC indicates that the SynRescue proteins do not form long-lived monomeric structures but instead form extended dimers that are kinetically unstable on the timescale of the chromatography experiments.

The SynRescue proteins are members of a third-generation library of binary patterned sequences designed to form α-helical bundles. The crystal structure of another protein from this same library, WA20, was solved recently and shown to form two extended α-helical hairpins, which intertwine to form a domain-swapped dimer (Fig. 1d) [22]. The sequences of the SynRescue proteins are 31–52% identical with WA20 and they behave similarly in CD, NMR, and SEC. Therefore, we suggest that the transient dimeric structures observed for the SynRescue proteins resemble the extended dimer seen in the X-ray crystal structure of WA20 (Fig. 1c) or a related structure with a different arrangement of helices similar to the models produced by Rosetta's fold-and-dock structure prediction protocol (Supplemental Fig. 3), or perhaps they sample a range of these structures as monomers and dimers.

Given the tendency of the third-generation sequences to sample dimeric states, we wished to understand which features in the design of the third-generation library promote this dimerization. We were particularly curious about this because the design of the third-generation library was inspired by the sequence of S824 (from a second-generation library), which formed a well-ordered monomeric 4-helix bundle with a disperse $^{1}H^{15}N$ HSQC NMR spectrum and a persistent structure that was readily solved by NMR [4].

We have identified three features that may have favored the formation of extended (double-length) α-helical hairpins that assemble into domain-swapped dimers. In each case, "negative design" might have prevented extension of the helices and the resulting dimerization [1]. These three features of negative design are summarized as follows:

(i) Breaking the hydrophobic register: The underlying premise of the binary patterning strategy is that matching the sequence periodicity of polar and nonpolar residues with the structural periodicity of the desired secondary structure will direct a chain to form amphiphilic secondary structures that bury hydrophobic side chains in the protein core. For α-helices, this requires placing nonpolar residues every 3 or 4 positions to match the helical repeat of 3.6 residues per turn. If this periodicity continues throughout a designed sequence, then one might expect the entire sequence to form one long amphiphilic helix. In particular, if the last nonpolar residue of one helix and the first nonpolar residue of the next helix are 3, 4, or 7 residues apart, then the two helices may form a single long helix with a continuous hydrophobic face. To avoid this possibility, one can use negative design to break this periodicity, offset the hydrophobic face of the helix, and disfavor the continuation of long helices. This feature of negative design was *not* incorporated into the third--generation library: thus, the sequences of the SynRescue proteins and WA20 have 7 residues from the last nonpolar residue of helix 1 (Trp23) to the first nonpolar residue of helix 2 (Leu30). Likewise, there are 7 residues from the last nonpolar residue of helix 3 (Leu75) to the first nonpolar residue of helix 4 (Val82). Since these sequences do not offset the hydrophobic register of an idealized amphiphilic α-helix, perhaps it is not surprising that the crystal structure of WA20 shows that helices 1 and 2 and helices 3 and 4 form continuous double-length helices. We presume the SynRescue sequences form similar extended helices in their dimeric structures.

(ii) Preventing favorable buried polar interactions: Another premise of the binary patterning strategy is that polar residues avoid burial. Therefore, in our libraries, polar residues are used only in positions designed to be on the solvent-exposed faces of helices or in interhelical loops. However, if these loops do not form at the expected locations and the helices continue through the intended loop sequences, then some of these polar residues will be on the buried faces of the extended helices. This is observed in the crystal structure of the WA20 dimer. Moreover, as shown in Fig. 1c, the sequences that were designed to form loops between helices 1 and 2 and between helices 3 and 4 pack against one another in the domain-swapped dimer. In the structure of WA20, the burial of these polar residues is enabled by a favorable electrostatic interaction between His26 and Glu78. Similarly, all the SynRescue proteins

studied here have charged and/or hydrogen bonding groups at positions 26 and 78 that could be satisfied by the formation of extended dimer structures similar to WA20. These residues at positions 26 and 78 are shown in boldface in Fig. 1a. These favorable buried polar interactions, which presumably stabilize the dimeric structure, could be prevented by using negative design to place similar charges at these sites (e.g., K/R26 and K/R78).

(iii) Interrupting helix propensity: Another way to use negative design to prevent the helices from extending through the intended loops would be to include helix breaking residues in the loops. The control proteins, S23 and S824, contain two glycines in each of the relevant loops. Glycine is well known as a helix breaker, and the structure of S824 shows the intended loops at these locations. Thus, both S23 and S824 are monomeric. In contrast, protein WA20 has no glycines at these positions, and its crystal structure shows helices that continue through the intended loop sequences. Likewise, the SynRescue proteins rarely have glycines in these regions and are presumed to form extended dimers similar to WA20.

While the features described above may have caused the SynRescue proteins to adopt less ordered structures, which vacillate between monomeric and dimeric states, this diminished order has *not* prevented the possibility of biological function. Quite the contrary, more than 20 different sequences from the third-generation library provide life-sustaining activities in *E. coli*: these sequences enable cell growth in strains that cannot grow in their absence [16]. These findings demonstrate that a well-ordered structure is *not* a prerequisite for biological function.

For natural proteins, structural biologists had long assumed that ordered structures are essential for biological function. However, this assumption arose, in part, from a bias that developed because the only protein structures that had been observed were those that "held still" long enough for their structures to be determined by crystallography or NMR. More recently, as new methods have been developed to study dynamic structures, it is becoming clear that many proteins essential for life are indeed dynamic and/or intrinsically disordered [14,15].

Advances in protein engineering provide additional compelling evidence that well-defined structures are not required for activity—even for high levels of enzyme catalysis. Most notably, Hilvert and coworkers demonstrated that an engineered version of chorismate mutase exists as a dynamic molten globule yet retains k_{cat} and K_m values similar to the wild-type enzyme [28].

Fluctuating or dynamic structures may have also played an important role in the early evolution of proteins. Jensen postulated that proteins did not have well-defined specific activities early in the history of life on earth. He suggested that primordial proteins had low levels of activity and low specificity. Instead of the highly specialized enzymes that we see in modern organisms, Jensen suggested that primordial proteins were promiscuous generalists. Broad specificity would have been advantageous at the early stages of molecular evolution because it would "maximize the catalytic versatility of an ancestral cell that functioned with limited enzyme resources" [29]. While Jensen's discussion of primordial proteins focused primarily on function, rather than structure, it seems reasonable to assume that nonspecific promiscuous functions would have been facilitated by nonspecific promiscuous structures. While it is not possible to go back in time to perform structural measurements and/or assay the biological fitness of primordial proteins, the *de novo* sequences in our libraries may in fact resemble the sequences that existed in the early history of life on earth.

Indeed, one of the *de novo* sequences described in the current study, SynIlvA1, has now been shown to be dynamic, in terms of both structure and function. The structural dynamics of SynIlvA1 are illustrated by the experiments described above, and recently, we reported the functional promiscuity of SynIlvA1, which was originally selected for its ability to rescue the isoleucine auxotroph Δ*ilvA* but also rescues Δ*fes*, which is essential for the assimilation of iron [30]. These observations suggest that dynamic proteins may not merely be "acceptable" structures for biological function but may in fact play key roles in evolutionary trajectories from multifunctional generalists to highly active specialists.

Methods

Computational simulations using Rosetta

Protein structure prediction simulations were performed using the Rosetta macromolecular modeling software fragment assembly protocol [24]. Briefly, this protocol combines 3-residue and 9-residue fragments (from high-resolution crystal structures) using a reduced centroid model of the protein, coarse-grained energy functions, and a Monte Carlo search procedure, followed by an all-atom high-resolution structure refinement step. The 3-residue and 9-residue fragments are chosen based on sequence similarity and predicted secondary structure of the target protein sequence. Fragments were generated using

the Robetta fragment server[†] and simulations were performed on a Princeton University Dell/SGI computer cluster with 10,304 cores. Sample command lines are given in the supplemental information.

To predict the structure of suspected oligomers, we used the Rosetta fold-and-dock protocol that has been used to predict the structure of protein oligomers [25]. We used the protocol to predict the structures of the proteins studied here under the assumption that they were symmetric homodimers with *C*2 symmetry. The fold-and-dock protocol essentially performs the standard Rosetta *ab initio* simulation while simultaneously docking monomers A and A′ in a symmetric complex, allowing translation and rotation in the x, y, and z directions. Sample command lines are given in the supplemental information.

Protein expression and purification

The genes for the proteins studied here are in a modified pCA24N vector [16]. The vector contains the chloramphenicol resistance gene [chloramphenicol acetyl transferase], an IPTG-inducible T5 promoter, and a ribosome binding site upstream from the gene of interest. The gene of interest is between a 5′ Nde1 site at the initiator methionine and a 3′ BsrG1 site that cleaves in the last four amino acids followed by a stop codon. Amino acid sequences for the constructs S824, S23, WA20, SynIlvA1, SynFes2, SynGltA1, SynSerB1, SynSerB2, SynSerB3, and SynSerB4 are listed in the supplemental information and with their European Nucleotide Archive accession number.

Proteins were expressed in *E. coli* BL21 (DE3) pLysS cells. Cells were grown in 1 L LB with 30 μg/mL chloramphenicol at 37 °C to an OD_{600} between 0.4 and 0.6 and were induced with 100 μM IPTG for 12–16 h at 18 °C. Cells were recovered by centrifugation at 5000*g* for 30 min. Cell pellets were resuspended in 50 mM sodium phosphate with 200 mM sodium chloride (pH 7.4) and were lysed by passing through an Emulsiflex C3 homogenizer at 15,000 psi for three cycles. Cell lysates were clarified by centrifugation at 7000*g* for 30 min. The supernatant was filtered using 0.22-μm PES membrane syringe filters.

Proteins were purified using immobilized metal affinity chromatography (IMAC). While our constructs do not contain a canonical histidine tag, they do contain a high percentage of histidines, on average 15%, and are readily purified using IMAC with a modified buffer system. The running buffer does not contain imidazole and is 50 mM sodium phosphate and 200 mM sodium chloride at pH 7.4 and the elution buffer is 50 mM sodium phosphate, 200 mM sodium chloride, and 500 mM imidazole at pH 7.4. The IMAC purification was performed as follows: filtered supernatant was applied to a 5-mL HisTRAP column (GE Healthcare) equilibrated in running buffer without imidazole. The column was washed with 5 column volumes of running buffer. A second wash step of 5 column volumes with 10% elution buffer removes proteins nonspecifically bound to the column, with the primary contaminating protein being chloramphenicol acetyl transferase. The proteins of interest were then eluted using 75% elution buffer. Eluted fractions were pooled, typically 10 mL, and further purified by SEC on a HiLoad Superdex 75 26/600 column (GE Healthcare). Purity of proteins from this two-step procedure was >95% as assessed by SDS-PAGE (Supplemental Fig. 6).

The proteins S824, S23, WA20, SynIlvA1, SynFes2, and SynSerB1 expressed and purified in high yield giving >30 mg/L expression culture. SynGltA1 and SynSerB2 expressed well but purified with lower yield giving ~10 mg/L of culture. SynSerB3 did not express at significant levels and was difficult to purify giving ~1 mg/L of culture. SynSerB4 had modest expression and did not purify in high yield giving ~5 mg/L of culture. It is interesting that SynSerB1 and SynSerB3 behaved so differently given that their amino acid sequences are 94% identical, with only six contiguous residues being different (Fig. 1). Additionally, the SynRescue proteins were prone to precipitation at protein concentrations above 200 μM, especially at sodium chloride concentrations below 100 mM.

CD spectroscopy

CD data were collected on a Chirascan CD spectrometer (Applied Photophysics). Far-UV CD spectra were collected using a 1-mm pathlength cuvette and protein concentrations of ~30 μM in 50 mM sodium phosphate and 100 mM sodium chloride at pH 7.4. Thermal denaturation experiments were performed by monitoring the α-helical CD signal at 222 nm, as the temperature was increased/decreased at 1 °C/min from 5 °C to 95 °C and then back to 5 °C. Thermal denaturation curves were fit to a two-state model of unfolding using gnuplot (see supplemental information for details).

NMR spectroscopy

NMR spectra were collected on an 800-MHz AVANCE III HD spectrometer (Bruker) with a 5-mm cryoprobe. Proteins were in 90% H_2O/10% D_2O with 50 mM sodium phosphate and 200 mM sodium chloride (pH 6.8). One-dimensional proton spectra were collected using WATERGATE solvent suppression [31]. Two-dimensional $^{1}H^{15}N$ HSQC were collected on uniformly labeled ^{15}N samples using the "hsqcfpf3gpphwg" pulse sequence from the Bruker library modified to use excitation sculpting water suppression. Labeled samples were grown as described previously, except that cultures were centrifuged and transferred to a minimal media containing 1.0 g/L of ^{15}N ammonium chloride prior to induction. NMR samples had concentrations ≥200 μM, as protein solubility allowed. All spectra were processed and visualized using TopSpin (Bruker) and CCPNMR [32].

SEC experiments

A Superdex 75 5/150 column (GE Healthcare) was used for analytical SEC. A set of standard proteins of BSA (66 kDa), carbonic anhydrase (29 kDa), cytochrome *c* horse heart (12.4 kDa), and aprotinin (6.5 kDa) were run on the column to measure elution volume, resolution, and sensitivity. Blue dextran (~2000 kDa) was used to identify the column void volume. These data were used to generate a standard curve of the ratio of elution volume over void volume *versus* Log10(molecular weight). The same was performed for the Superdex 75 10/300 and Superdex 75 26/600 columns. The SEC experiments were performed using the same samples concentrated for the

$^{1}H^{15}N$ HSQC experiments, with concentrations of ≥200 μM and also at dilutions of 30 μM both in 50 mM sodium phosphate and 200 mM sodium chloride at pH 6.8 giving similar results. The injection volumes were 1000 μL, 500 μL, and 100 μL for the 26/600, the 10/300, and the 5/150 columns. The flow rates were 2.6 mL/min, 1.0 mL/min, and 0.5 mL/min for the 26/600, the 10/300, and the 5/150 columns. Molecular weights were calculated from elution volumes by rearranging the standard curve equation for the S75 5/150 to be MW = 10^(−2.1362*(EV/3.0)/0.48 + 7.3678), S75 10/300 to be MW = 10^(−1.5068*(EV/24)/0.32 + 6.5893), and S75 26/600 to be MW = 10^(−1.0806* (EV/318)/0.37 + 6.0493).

Acknowledgements

We thank Dr. Istvan Pelczer and Ken Conover from the Princeton University Chemistry Department NMR facility for helpful discussions on NMR pulse sequences and results. We also thank the Princeton University Research Computing center for access to the Tiger cluster. We also thank Ann Mularz and Katherine Digianantonio for helpful discussions on this research and manuscript. This work was funded by National Science Foundation grants MCB-1050510 and MCB-1409402 to M.H.H. and a National Institutes of Health F32 fellowship (1F32GM106622) to G.S.M.

Author Contributions: G.S.M., J.B.G., and M.H.H. designed the research. G.S.M. and J.B.G. performed the experiments. G.S.M., J.B.G., and M.H.H. analyzed the data. G.S.M., J.B.G., and M.H.H. wrote the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2015.12.008.

Received 9 September 2015;
Received in revised form 15 December 2015;
Accepted 15 December 2015
Available online 18 December 2015

Keywords:
de novo protein design;
helix bundle;
synthetic biology;
artificial proteomes

Present address: J. B. Greisman, D. E. Shaw Research, New York, NY 10036, USA.

†http://robetta.bakerlab.org/.

Abbreviations used:
NOE, Nuclear Overhauser effect; HSQC, heteronuclear single quantum coherence; SEC, size-exclusion chromatography; IMAC, immobilized metal affinity chromatography.

References

[1] M.H. Hecht, J.S. Richardson, D.C. Richardson, R.C. Ogden, *De novo* design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence, Science 249 (1990) 884–891.
[2] L. Regan, W.F. DeGrado, Characterization of a helical protein designed from first principles, Science 241 (1988) 976–978.
[3] P.B. Harbury, J.J. Plecs, B. Tidor, T. Alber, P.S. Kim, High-resolution protein design with backbone freedom, Science 282 (1998) 1462–1467.
[4] Y. Wei, S. Kim, D. Fela, J. Baum, M.H. Hecht, Solution structure of a *de novo* protein from a designed combinatorial library, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 13270–13273.
[5] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy, Science 302 (2003) 1364–1368.
[6] G.S. Murphy, B. Sathyamoorthy, B.S. Der, M.C. Machius, S.V. Pulavarti, T. Szyperski, B. Kuhlman, Computational *de novo* design of a four-helix bundle protein—DND_4HB, Protein Sci. 24 (2015) 434–445.
[7] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T.B. Acton, G.T. Montelione, D. Baker, Principles for designing ideal protein structures, Nature 491 (2012) 222–227.
[8] P.S. Huang, G. Oberdorfer, C. Xu, X.Y. Pei, B.L. Nannenga, J.M. Rogers, F. DiMaio, T. Gonen, B. Luisi, D. Baker, High thermodynamic stability of parametrically designed helical bundles, Science 346 (2014) 481–485.
[9] S.J. Fleishman, T.A. Whitehead, D.C. Ekiert, C. Dreyfus, J.E. Corn, E.M. Strauch, I.A. Wilson, D. Baker, Computational design of proteins targeting the conserved stem region of influenza hemagglutinin, Science 332 (2011) 816–821.
[10] D. Rothlisberger, O. Khersonsky, A.M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J.L. Gallaher, E.A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K.N. Houk, D.S. Tawfik, D. Baker, Kemp elimination catalysts by computational enzyme design, Nature 453 (2008) 190–195.
[11] J.B. Siegel, A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert, J.L. St Clair, J.L. Gallaher, D. Hilvert, M.H. Gelb, B.L. Stoddard, K.N. Houk, F.E. Michael, D. Baker, Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, Science 329 (2010) 309–313.
[12] L. Jiang, E.A. Althoff, F.R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J.L. Gallaher, J.L. Betker, F. Tanaka, C.F. Barbas III, D. Hilvert, K.N. Houk, B.L. Stoddard, D. Baker, *De novo* computational design of retro-aldol enzymes, Science 319 (2008) 1387–1391.
[13] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, D.C. Phillips, A three-dimensional model of the myoglobin molecule obtained by X-ray analysis, Nature 181 (1958) 662–666.
[14] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.

[15] C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins and intrinsically disordered protein regions, Annu. Rev. Biochem. 83 (2014) 553–584.
[16] M.A. Fisher, K.L. McKinley, L.H. Bradley, S.R. Viola, M.H. Hecht, *De novo* designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth, PLoS One 6 (2011) e15364.
[17] L.H. Bradley, R.E. Kleiner, A.F. Wang, M.H. Hecht, D.W. Wood, An intein-based genetic selection allows the construction of a high-quality library of binary patterned *de novo* protein sequences, Protein Eng. Des. Sel. 18 (2005) 201–207.
[18] S.C. Patel, L.H. Bradley, S.P. Jinadasa, M.H. Hecht, Cofactor binding and enzymatic activity in an unevolved superfamily of *de novo* designed 4-helix bundle proteins, Protein Sci. 18 (2009) 1388–1400.
[19] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M.H. Hecht, Protein design by binary patterning of polar and nonpolar amino acids, Science 262 (1993) 1680–1685.
[20] Y. Wei, T. Liu, S.L. Sazinsky, D.A. Moffet, I. Pelczer, M.H. Hecht, Stably folded *de novo* proteins from a designed combinatorial library, Protein Sci. 12 (2003) 92–102.
[21] A. Go, S. Kim, J. Baum, M.H. Hecht, Structure and dynamics of *de novo* proteins from a designed superfamily of 4-helix bundles, Protein Sci. 17 (2008) 821–832.
[22] R. Arai, N. Kobayashi, A. Kimura, T. Sato, K. Matsuo, A.F. Wang, J.M. Platt, L.H. Bradley, M.H. Hecht, Domain-swapped dimeric structure of a stable and functional *de novo* four-helix bundle protein, WA20, J. Phys. Chem. B 116 (2012) 6789–6797.
[23] I. Cherny, M. Korolev, A.N. Koehler, M.H. Hecht, Proteins from an unevolved library of *de novo* designed sequences bind a range of small molecules, ACS Synth. Biol. 1 (2012) 130–138.
[24] P. Bradley, K.M. Misura, D. Baker, Toward high-resolution *de novo* structure prediction for small proteins, Science 309 (2005) 1868–1871.
[25] R. Das, I. Andre, Y. Shen, Y. Wu, A. Lemak, S. Bansal, C.H. Arrowsmith, T. Szyperski, D. Baker, Simultaneous prediction of protein folding and docking at high resolution, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 18978–18983.
[26] S.Y. Lau, A.K. Taneja, R.S. Hodges, Synthesis of a model protein of defined secondary and quaternary structure. Effect of chain length on the stabilization and formation of two-stranded alpha-helical coiled-coils, J. Biol. Chem. 259 (1984) 13253–13261.
[27] H.P. Erickson, Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy, Biol. Proced. Online 11 (2009) 32–51.
[28] K. Vamvaca, B. Vogeli, P. Kast, K. Pervushin, D. Hilvert, An enzymatic molten globule: Efficient coupling of folding and catalysis, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 12860–12864.
[29] R.A. Jensen, Enzyme recruitment in evolution of new function, Annu. Rev. Microbiol. 30 (1976) 409–425.
[30] B.A. Smith, A.E. Mularz, M.H. Hecht, Divergent evolution of a bifunctional *de novo* protein, Protein Sci. 24 (2015) 246–252.
[31] M. Piotto, V. Saudek, V. Sklenar, Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions, J. Biomol. NMR 2 (1992) 661–665.
[32] W.F. Vranken, W. Boucher, T.J. Stevens, R.H. Fogh, A. Pajon, M. Llinas, E.L. Ulrich, J.L. Markley, J. Ionides, E.D. Laue, The CCPN data model for NMR spectroscopy: Development of a software pipeline, Proteins 59 (2005) 687–696.