JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

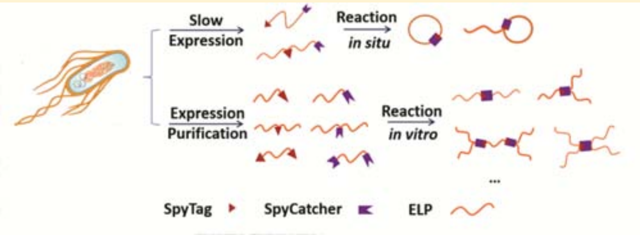
pubs.acs.org/JACS

Controlling Macromolecular Topology with Genetically Encoded SpyTag–SpyCatcher Chemistry

Wen-Bin Zhang,[†] Fei Sun,[†] David A. Tirrell,* and Frances H. Arnold*

Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States

Supporting Information



ABSTRACT: Control of molecular topology constitutes a fundamental challenge in macromolecular chemistry. Here we describe the synthesis and characterization of artificial elastin-like proteins (ELPs) with unconventional nonlinear topologies including circular, tadpole, star, and H-shaped proteins using genetically encoded SpyTag–SpyCatcher chemistry. SpyTag is a short polypeptide that binds its protein partner SpyCatcher and forms isopeptide bonds under physiological conditions. Sequences encoding SpyTag and SpyCatcher can be strategically placed into ELP genes to direct post-translational topological modification *in situ*. Placement of SpyTag at the N-terminus and SpyCatcher at the C-terminus directs formation of circular ELPs. Induction of expression at 16 °C with 10 μM IPTG yields 80% monomeric cyclic protein. When SpyTag is placed in the middle of the chain, it exhibits an even stronger tendency toward cyclization, yielding up to 94% monomeric tadpole proteins. Telechelic ELPs containing either SpyTag or SpyCatcher can be expressed, purified, and then coupled spontaneously upon mixing *in vitro*. Block proteins, 3-arm or 4-arm star proteins, and H-shaped proteins have been prepared, with the folded CnaB2 domain that results from the SpyTag–SpyCatcher reaction as the molecular core or branch junction. The modular character of the SpyTag–SpyCatcher strategy should make it useful for preparing nonlinear macromolecules of diverse sequence and structure.

■ INTRODUCTION

The most fundamental challenge in macromolecular synthesis is the creation of tailor-made materials with complete control of the molecular framework, comparable to what can be achieved in small molecules through synthetic organic chemistry.[1] Although chemical polymerization processes provide only limited control of chain length, sequence, and stereochemistry, they can be engineered to yield remarkably diverse topologies, including linear and cyclic chains, branched structures, stars, and networks.[2,3] In contrast, the translational machinery of cells provides exquisite control of the size, sequence, and folded structure of each cellular protein but is limited, for the most part, to the synthesis of linear chains. It is of interest both academically and technologically to combine the precise structural control provided by protein synthesis with the topological variations characteristic of synthetic polymers to develop new biomaterials that exhibit diverse biological and physical properties.

The lack of topological diversity in natural proteins is compensated to a large extent by protein folding, which directs the placement of atoms in space and defines their biological function. The nonlinear topologies that do exist result most frequently from post-translational modifications. For example, naturally occurring circular proteins have attracted considerable attention owing to their prevalence, sequence diversity, and unique functions.[4,5] They are derived from longer precursor proteins through extensive processing events, although the detailed mechanisms of cyclization remain largely unknown.[4−6] Subunits that make up the bacteriophage HK97 viral capsid have been found to arrange into topologically linked "chain-mail" structures through side-chain-mediated covalent bond formation that enhances capsid stability.[7] Chain-extended proteins, including branched structures,[8] have been observed in polyubiquitination, which is central to many biological functions.[9,10] There has been no systematic study, however, of the variation in protein topology that might be achieved through protein engineering, despite substantial progress in the development of efficient protein–protein ligation methods. Native chemical ligation,[11,12] subtiligase-catalyzed cyclization,[13] disulfide preorganization,[14] and other standard amide bond-forming reactions as well as "click" reactions[15] have all been used to cyclize unstructured polypeptides. For folded proteins, the distance between the termini is critical in determining the efficiency of cyclization. Techniques such as split-intein technology[16−18] and sortase-mediated cyclization[19] have been used to cyclize folded proteins including β-lactamase,[18] green fluorescent protein,[17,19] and dihydrofolate reductase with varying degrees of success.[16] However, these methods cannot be adapted easily to prepare other protein topologies such as domain-selective cyclized proteins (tadpole-like proteins) or star-like structures. Here we describe a versatile, modular biosynthetic route to branched and cyclic macromolecules of well-defined structure.

Received: July 24, 2013
Published: August 21, 2013

Recently, Howarth and co-workers developed a genetically encodable, highly reactive peptide (SpyTag, 1.1 kDa)−protein (SpyCatcher, 12 kDa) pair by splitting the autocatalytic isopeptide bond-forming subunit (CnaB2 domain) of *Streptococcus pyogenes*.[20,21] Upon mixing, SpyTag and SpyCatcher undergo autocatalytic isopeptide bond formation between Asp^{117} on SpyTag and Lys^{31} on SpyCatcher. The reaction is compatible with the cellular environment and highly specific for protein/peptide conjugation. Since the reactive units are conveniently introduced by genetic engineering, we recognized that the SpyTag−SpyCatcher chemistry is ideal for preparing proteins with different topologies. We envisioned that strategic placement of the sequences encoding SpyTag and SpyCatcher within protein-coding genes would program the post-translational modification of the expressed proteins *in situ* and enable the synthesis of unconventional protein topologies such as circular proteins and tadpole-like proteins. Alternatively, telechelic proteins could be prepared with reactive groups at defined locations within the protein chain, purified, and subsequently coupled *in vitro* to yield branched and cyclic structures.

Here we report our efforts to synthesize proteins with complex architectures including cyclic, tadpole-like, star, and branched topologies using the SpyTag−SpyCatcher chemistry. We demonstrate these ideas by modifying artificial elastin-like proteins (ELPs) similar to the elastins found in connective tissues and widely used as model extracellular matrices.[22] The ELP sequence chosen for this study is based on the hydrophilic polypentapeptide $(VPGXG)_n$, where X is a mixture of glutamic acid and valine.[23] In nature, extensive processing, including cross-linking of the linear elastin precursor, is required to impart elasticity and resilience to the tissue. The results described here provide a basis for the study of topological effects on the material properties and self-assembly behavior of ELPs and other engineered proteins.[24]

EXPERIMENTAL SECTION

DNA Construction. The ELP gene was ordered from Genscript. The SpyTag coding sequence was obtained from IDT (Integrated DNA Technologies) as short complementary oligonucleotides and subsequently annealed. The SpyCatcher sequence was purchased as a gBlocks gene fragment from IDT and amplified by the polymerase chain reaction (PCR). The PCR primers included the restriction sites needed for cloning. The coding sequences were cloned into the bacterial expression vector pQE-80L (Qiagen Inc.) in the correct order by standard restriction digestion and ligation protocols to give the plasmids containing the open reading frames shown in Figure 1. For the AB20D construct, a control construct, AB20A, was prepared by QuickChange mutagenesis on AB20D using the primers 5′-CGTCGACGCCCATATTGTCATGGTTGCTGCATACAAGCCGACGAAGCTCGACGGCCAC-3′ and 5′-GTGGCCGTCGAGCTTCGTCGGCTTGTATGCAGCAACCATGACAATATGGGCGTCGACG-3′ based on the manufacturer's recommended protocol (Stratagene Inc.). The sequences of all genes were verified by direct DNA sequencing. The plasmids used in the paper are summarized in the Supporting Information (Table S1).

Protein Synthesis and Purification. Plasmids were transformed into chemically competent *Escherichia coli* strain BL21 for expression. A single colony was inoculated into 5 mL of 2XYT broth containing 100 μg/mL ampicillin and incubated overnight in a shaker at 37 °C. The overnight culture was inoculated into 1 L of 2XYT broth containing 100 μg/mL ampicillin and grown at 37 °C with vigorous shaking until the OD_{600} reached 0.5−0.7. The shaker was then adjusted to the designated expression temperature. For leaky expression, no inducer was added, and shaking was continued for another 8 h at 37 °C or 12 h at 16 °C. For induced expression, isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added to the culture. After the designated induction time, the cells were harvested. For circular proteins and tadpole proteins, the cells were lysed under denaturing conditions (in 8 M urea) so as to quench further reactions that could alter the product distribution. For telechelic proteins containing the SpyCatcher domain (such as EB, EBE, and BB), the cells were lysed under native conditions. For telechelic proteins containing the SpyTag motif (such as EA, EAE, and AA), the cells could be lysed either under native conditions or under denaturing conditions. Purification was performed as described in the Qiagen Expressionist. For purification under native conditions, the cleared lysate was mixed with a 50% Ni-NTA slurry and agitated gently at 4 °C for 1 h. The mixture was then loaded into a column, washed with wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH = 8.0), and then eluted with elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, pH = 8.0). For purification under denaturing conditions, the protocol was the same, except different buffers were used: lysis buffer (100 mM NaH_2PO_4, 10 mM Tris·Cl, 8 M urea, 10 mM imidazole, pH = 8.0), wash buffer (100 mM NaH_2PO_4, 10 mM Tris·Cl, 8 M urea, 20 mM imidazole, pH = 8.0), and elution buffer (100 mM NaH_2PO_4, 10 mM Tris·Cl, 8 M urea, 20 mM imidazole, pH = 4.5). After purification, the proteins were dialyzed against ddH_2O and lyophilized to give white powders. For circular protein and tadpole proteins, yields varied from 5 to 20 mg/L depending on the expression conditions. For homotelechelic proteins, the yields were approximately 25−40 mg/L.

Protein Coupling. Proteins were dissolved in 50 mM phosphate buffered saline (PBS) at pH = 7.5 at a concentration of 20 μM. Coupling was performed by mixing SpyTag and SpyCatcher solutions in appropriate stoichiometric ratios.

Protein Characterization. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was performed to determine the apparent molecular weight of each protein. Chain-extended products and small differences between protein topological isomers were analyzed by size exclusion chromatography with a Superdex 200 10/300GL column in an ATKA FPLC system (GE Healthcare). The buffer was PBS (pH = 7.5); the flow rate was 0.5 mL/min. Matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) was conducted on an Applied Biosystems Voyager mass spectrometer with sinapinic acid as the matrix. Proteolytic digestion was performed with ProTEV Plus (Promega Inc.) according to the general protocol given by the manufacturer. To 460 μL of protein solution were added 25 μL of ProTEV buffer (20×), 5 μL of 100 mM DTT, and 10 μL of ProTEV Plus. The mixture was then left at room temperature overnight.

RESULTS AND DISCUSSION

General Design. The telechelic ELP gene constructs designed and used in this work are shown in Figure 1. We examined two types of telechelic ELPs: (1) heterotelechic polymers that contain both SpyTag and SpyCatcher (Figure 1a,c) and (2) homotelechelics that contain either SpyTag or SpyCatcher (Figure 1d−i). In the former, the placement of the sequences encoding SpyTag and SpyCatcher within the protein-coding genes programs the post-translational modification of the expressed proteins *in situ*. For example, placing SpyTag and SpyCatcher at the N- and C-termini (Figure 1a, AB20D), respectively, of ELPs should lead to cyclized proteins (Scheme 1). The reactive amino acid residues (20D and 255K) of AB20D are shown in Figure 1a. If SpyTag is placed in the middle of the chain and SpyCatcher at the C-terminus (Figure 1b, EAEB), only the domain flanked by SpyTag and SpyCatcher will be cyclized, leading to tadpole-like proteins (Scheme 2). In both constructs, a tobacco etch virus (TEV) protease digestion site was placed immediately N-terminal to the SpyCatcher block to allow determination of product topology via proteolytic digestion. We also prepared control

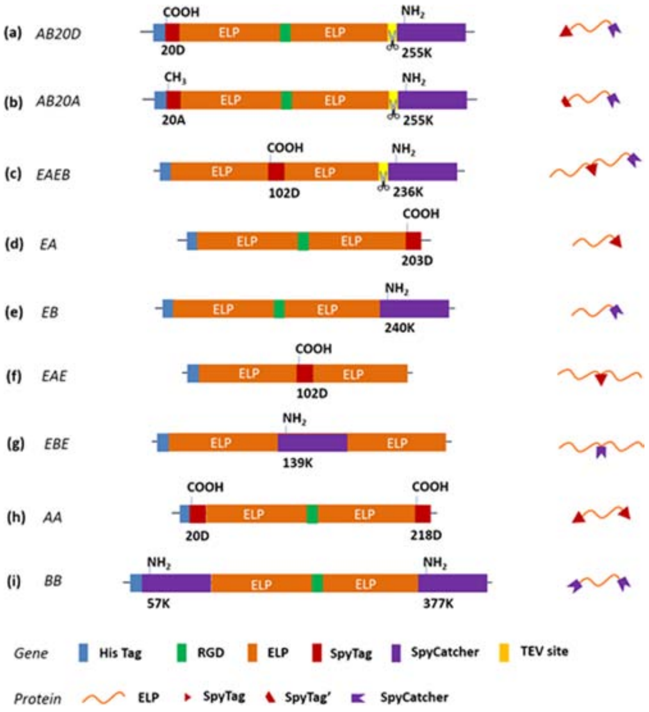
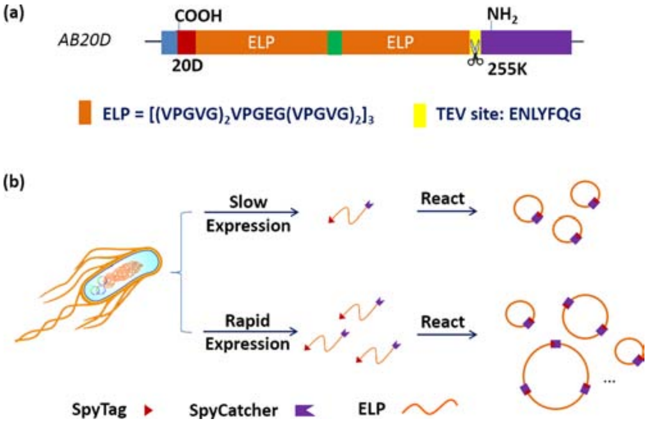


Figure 1. Gene constructs and the corresponding telechelic proteins: CBD, cell-binding domain; ELP, elastin-like protein; SpyTag′, SpyTag with D117A mutation that abolishes its reactivity; TEV site, ENLYFQG sequence that will be recognized and digested by TEV protease. The RGD cell-binding domain was included in some constructs to enable future application as model extracellular matrix proteins.

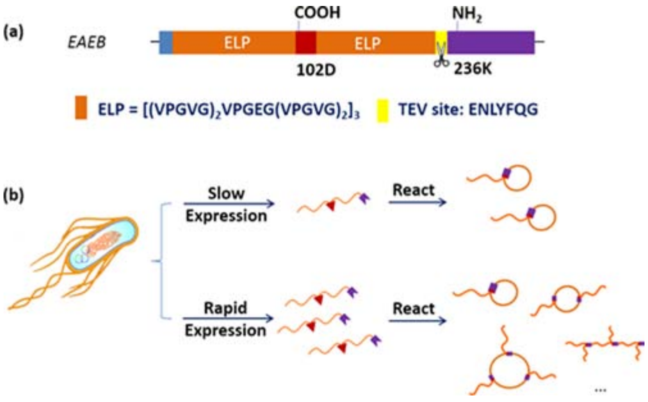
Scheme 1. (a) Schematic Illustration of Construct Encoding AB20D: Sequences of ELP and the TEV Digestion Site; (b) Slow Expression in *E. coli* Leads to *in Situ* Cyclization of the Protein, whereas Rapid Expression Leads to Significant Amounts of Chain-Extended Products



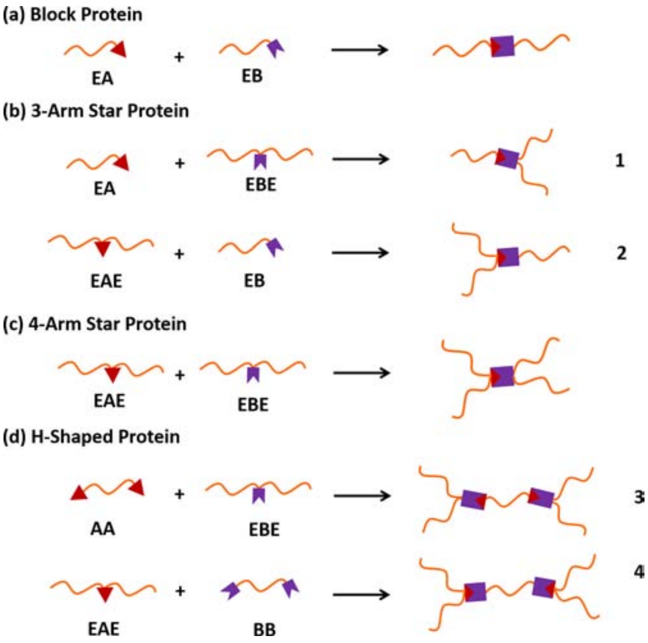
mutant *AB20A*, in which the reactive aspartic acid residue was mutated to a nonreactive alanine. The mutation is expected to abolish covalent bond formation while leaving molecular recognition and binding between SpyTag and SpyCatcher unaffected.[21]

The second class of telechelic proteins contains one or more reactive units at terminal or internal positions (Figure 1d–i). We expressed and purified each of these proteins separately. Reactions between these proteins are expected to give various star- or H-shaped proteins (Scheme 3). The number and position of the reactive units can be varied widely; only the simplest examples are illustrated here to provide proof-of-concept. The amino acid sequences for all these proteins are given in the Supporting Information (Figures S2, S6, S8, S10, and S12).

Scheme 2. (a) Schematic Illustration of Construct Encoding EAEB: Sequences of ELP and the TEV Digestion Site; (b) Slow Expression Leads to *in Situ* Domain-Selective Cyclization, Giving Tadpole-like Proteins, whereas Rapid Expression Leads to Chain-Extended Products in Addition to Tadpole-like Proteins



Scheme 3. Topological Diversification *in Vitro* to (a) Block Protein, (b) 3-Arm Star Proteins 1 and 2, (c) a 4-Arm Star Protein, and (d) H-Shaped Proteins 3 and 4



Cyclization of ELP in Living Cells. The synthesis of cyclic macromolecules is challenging, both for proteins and for synthetic polymers.[25] In polymer chemistry, a telechelic polymer with two mutually reactive groups at the chain ends can be regarded as an AB-type monomer that undergoes two types of reactions, cyclization and/or chain extension, depending on concentration. Cyclization is often achieved in extremely dilute solutions where chain extension is slow. To improve the synthesis of cyclic polymers, several ingenious methods have been developed, including ring-expansion metathesis,[26] electro-

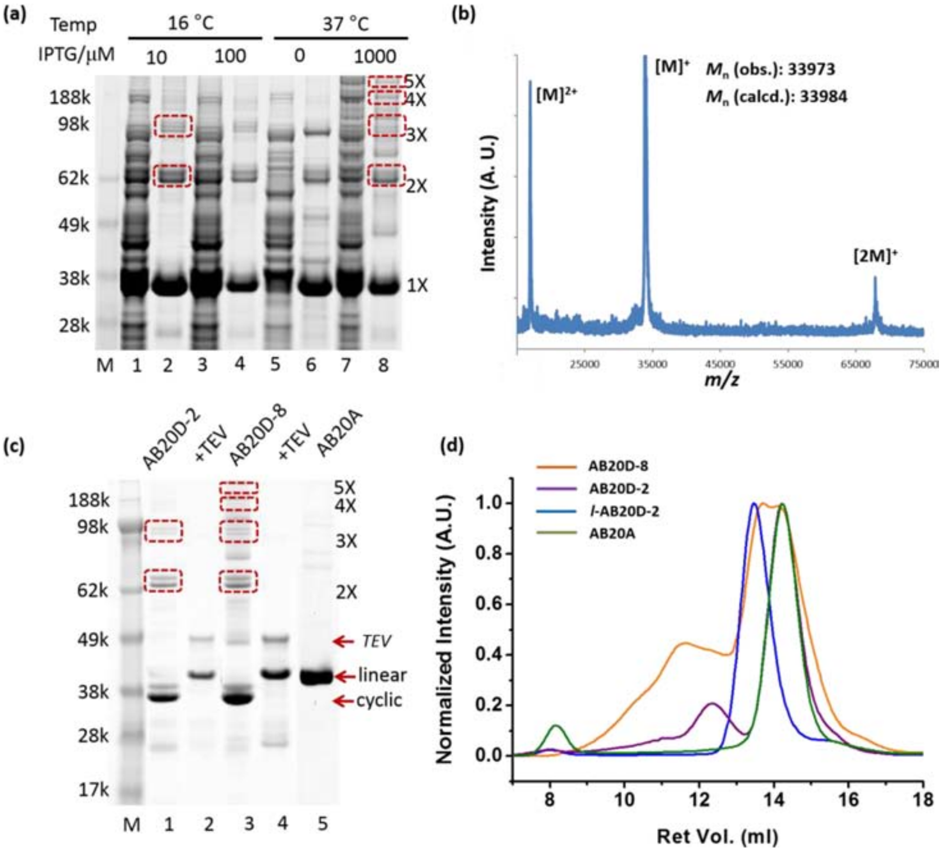


Figure 2. (a) SDS-PAGE analysis of AB20D expression under different conditions. Lane M is the MW marker. Lanes 1−4 show samples expressed at 16 °C; the samples in lanes 5−8 were expressed at 37 °C. Lanes 1, 3, 5, and 7 are lysates; lanes 2, 4, 6, and 8 are the corresponding purified products and are designated AB20D-2, -4, -6, and -8, respectively. The concentration of added inducer (isopropyl-1-β-D-thiogalactoside (IPTG)) is indicated on top of each lane. (b) MALDI-TOF mass spectrum of the cyclic product obtained from expression at 16 °C with 0.01 mM IPTG (AB20D-2). (c) SDS-PAGE analysis of proteolytic digestion products and the linear control AB20A. Lane M is the MW marker. Lanes 1 and 3 are AB20D-2 and AB20D-8 before digestion. Lanes 2 and 4 are the corresponding products after digestion. The band with an apparent molecular weight of 48k is the TEV protease. Lane 5 is the linear control mutant AB20A. (d) Overlay of SEC traces of AB20D-2 (purple curve), AB20D-8 (brown curve), relinearized AB20D obtained by proteolytic digestion of AB20D-2 (*l*-AB20D-2, blue curve), and linear control AB20A (green curve). The void volume of the column is approximately 8 mL.

static self-assembly and covalent fixation,[27] and "click"-reaction-assisted cyclization.[28] In the cell, an effectively infinitesimal concentration of monomeric telechelic protein can be achieved simply by reducing the protein synthesis rate. For a given bacterial expression host, the protein synthesis rate depends on temperature,[29,30] induction level (IPTG concentration),[31] and culture medium.[31] It has been shown that *E. coli* exhibits higher chain elongation rates at higher temperatures as well as a defect in initiation at low temperatures.[29] It has also been reported that protein expression levels increase with increasing IPTG concentration up to concentrations of ~100 μM.[31] We reasoned that slow expression of heterotelechelic proteins such as AB20D at low temperature and/or without induction would lead predominantly to *in situ* protein cyclization, whereas expression at 37 °C with full induction would promote chain extension.

To evaluate this hypothesis, plasmid pQE-*AB20D*, which carries the *AB20D* gene under control of the T5 promoter, was used to transform chemically competent *E. coli* strain BL-21. This plasmid allows leaky expression of the *AB20D* gene in the absence of IPTG. Expression was tested in 2XYT medium at 16 and 37 °C with varying concentrations of IPTG. Cells were harvested and lysed under denaturing conditions to prevent further reaction. The target protein was then purified by nickel-affinity chromatography and characterized by SDS-PAGE, MALDI-TOF mass spectrometry, and size exclusion chromatography (SEC). The topology of the product was examined by proteolytic digestion and compared with control samples. The results are summarized in Figure 2.

Figure 2a shows SDS-PAGE analysis of lysates and the corresponding purified proteins obtained from expression under different conditions. In lysates prepared from induced cultures (lanes 1, 3, and 7), overexpression is indicated by the appearance of a strong protein band just below the 38k MW marker. There is no such prominent band in the lysate prepared from the uninduced culture (lane 5). Analysis of the purified products (lanes 2, 4, 6, and 8) reveals a prominent band in a position just below the 38k MW marker in each sample. The MALDI-TOF mass spectra obtained on these purified samples (Figures 2b and S3) all show a major peak at m/z 33 973, within 0.1% of the expected value of 33 984. The product obtained at 37 °C contained significant amounts of chain-extended species, including dimer, trimer, and tetramer (Figure S3).

Because both SDS-PAGE and MALDI-TOF mass spectrometry are biased toward detection of low molecular weight (typically <200k) proteins, SEC was used to provide a more complete analysis of high MW products. The SEC overlay is

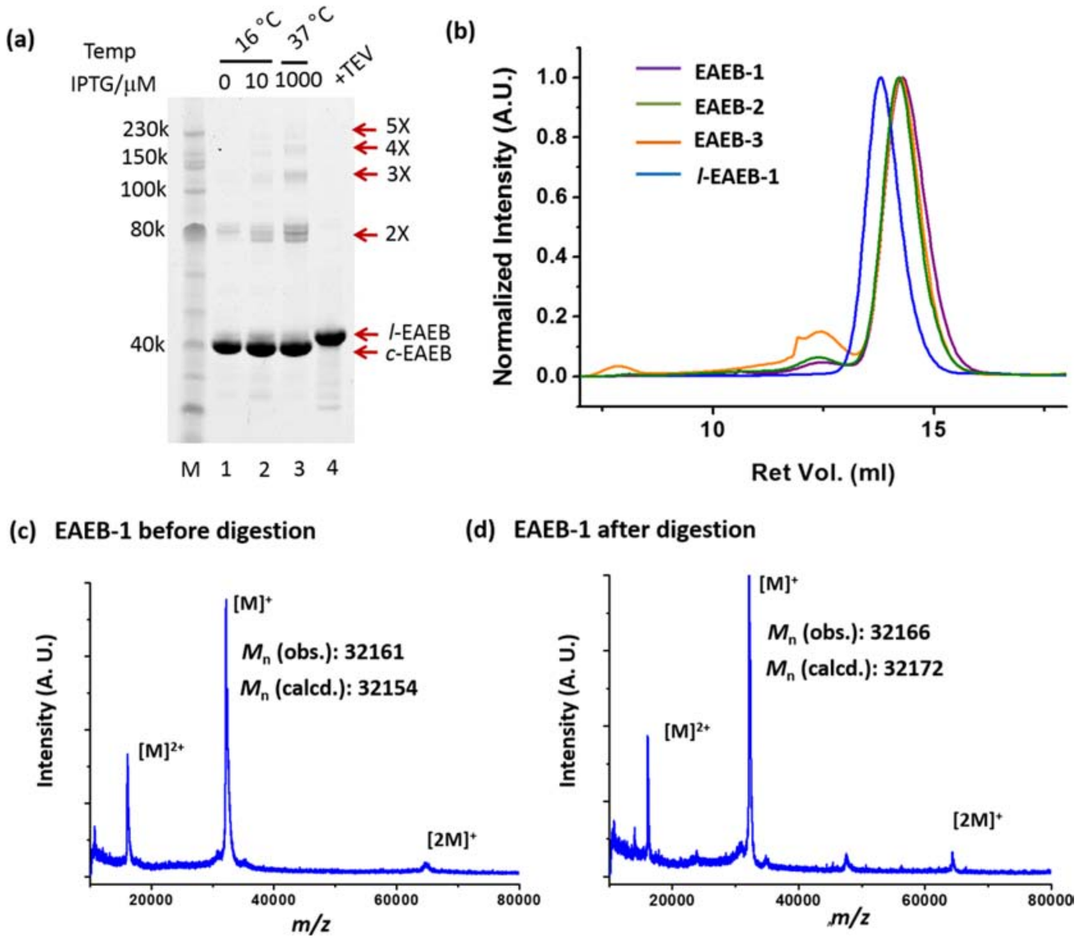


Figure 3. (a) SDS-PAGE analysis of purified EAEB proteins expressed under different conditions. Lane M is the MW marker. Lanes 1 and 2 show samples expressed at 16 °C with no induction and in the presence of 10 μM IPTG, respectively. Lanes 3 shows a sample expressed at 37 °C with 1 mM IPTG induction. The samples are designated EAEB-1, -2, and -3, respectively. Lane 4 is relinearized EAEB-1 obtained by proteolytic digestion (*l*-EAEB-1). (b) Overlay of SEC traces of EAEB-1 (purple curve), EAEB-2 (green curve), EAEB-3 (brown curve), and *l*-EAEB-1 (blue curve). (c) MALDI-TOF mass spectrum of EAEB-1. (d) MALDI-TOF mass spectrum of *l*-EAEB-1. The MW remains essentially unchanged upon digestion.

shown in Figure 2d. The chromatogram of AB20D-2 shows ~80% monomeric protein in addition to some chain-extended products, consistent with the SDS-PAGE results (Figure 2a, lane 2). By contrast, the chromatogram of AB20D-8 is multimodal and exhibits a distinctive tail that approaches the void volume of the column, suggesting that the longest chains in AB20D-8 are close to the exclusion limit (MW ≈ 1.3×10^6). Integration of the peak area indicates that the monomeric species constitute ~67% of the product; the remaining species are oligomers of varying degrees of chain extension. For example, a molecular weight of 1.3×10^6 corresponds to 38 AB20D monomers. Given that artificial proteins of such high MW are difficult to prepare by direct gene expression, polymerizing smaller protein units provides a useful new tool for macromolecular synthesis.[32]

The linear control sample was obtained by expression of the mutant protein AB20A. Because the mutation abolishes isopeptide bond formation between Asp20 and Lys255,[21] we anticipated that no covalent cyclization would occur in AB20A. Nevertheless, noncovalent binding between SpyTag and SpyCatcher persists, even in the mutant. As expected, only a single product was observed in SDS-PAGE analysis of AB20A (Figure 2c, lane 5). The MALDI-TOF mass spectrum of the purified product (Figure S4) shows a MW of 33960, which matches the calculated value of 33958. Despite the near-identical MWs, AB20A exhibits significantly lower mobility than the AB20D-2 and AB20D-8 monomers on SDS-PAGE (Figure 2c). This observation suggests that the monomeric species in AB20D-2 and AB20D-8 are cyclic, because cyclized proteins generally exhibit higher mobility than their linear counterparts on SDS-PAGE due to more compact conformations.[17] It also confirms that Asp20 is essential for covalent cyclization. In Figure 2c, the light bands in lanes 1 and 3, located slightly above the main cyclic monomer band but below the linear control band, are likely knotted protein polymers with structures more compact than that of the linear polymer.

To further verify that the monomeric species in AB20D-2 and AB20D-8 are cyclic, the proteins were digested with TEV protease. If the protein is linear, protease digestion should generate two linear fragments of MW 22k and 12k. If the protein is circular, digestion yields a single linear product with essentially the same MW as the precursor. When AB20D-2 was treated with TEV protease, the major band in SDS-PAGE shifted to significantly lower mobility (Figure 2c, lane 2). Meanwhile, the MALDI-TOF mass spectrum of the digested product (Figure S5) indicated a MW of 34 014, close to that expected for the full-length linear protein (34 002), suggesting that proteolysis changes only the protein topology. We thus conclude that the high-mobility bands observed prior to proteolysis are cyclic proteins and that the new band resulting from protease digestion is the relinearized product (*l*-AB20D). This conclusion is consistent with the SEC results (Figure 2d). The relinearized *l*-AB20D appears at lower retention volume than the monomeric precursor in AB20D-2. The lower retention volume suggests an increase in hydrodynamic volume upon digestion, which is common for relinearization of cyclic polymers.[18,25,33] Upon digestion, the oligomeric species in AB20D-2 and AB20D-8 (Figure 1a,c in red rectangles) were all converted to a single linear monomer. The appearance of multiple bands for each of the oligomers suggests that they are likely mixtures of cyclic forms and more complex topologies such as knotted cycles.

All of these results suggest that AB20D has a strong tendency to cyclize, even under overexpression conditions. What contributes to the high efficiency of cyclization? The SEC overlay of AB20D-2, *l*-AB20D-2, and AB20A sheds some light on this question (Figure 2d). Although AB20A and *l*-AB20D are both linear proteins, they appear at different retention volumes. The elution profile of AB20A overlaps with that of AB20D-2 (Figure 2d) except for the small oligomer peak and high MW tailing, suggesting that in solution both AB20A and the monomeric species in AB20D-2 adopt cyclic conformations. Thus, molecular recognition between SpyTag and SpyCatcher preorganizes the proteins into a cyclic conformation, which is subsequently covalently fixed in AB20D but not in AB20A. These results highlight the importance of molecular recognition in achieving highly specific and efficient cyclization.

Domain-Selective Cyclization to Tadpole Proteins. The method just described provides a versatile new strategy for the direct cellular synthesis of circular proteins. Yet it has advantages beyond the synthesis of circular proteins. Covalent closure of the amide backbone often confers useful properties, such as enhanced stability or improved bioactivity, on circular proteins and peptides.[6] For multidomain proteins, it is readily imagined that domain-selective cyclization could be advantageous in tailoring protein properties.[17,18,34] However, to the best of our knowledge, domain-selective cyclized proteins have not yet been found in nature, nor have they been explored in protein engineering. The established techniques for circular protein synthesis such as native chemical ligation,[11] split-intein technology,[16−18] and sortase-mediated cyclization[19] cannot be easily adapted to domain-selective cyclization. Since SpyTag is known to be reactive at internal sites as well as terminal sites,[21] the SpyTag−SpyCatcher chemistry is naturally suited for this purpose. If SpyTag is placed in the interior of the chain, only the domain that is flanked by SpyTag and SpyCatcher will be cyclized. Construct EAEB was used to examine this idea.

Plasmid pQE-*EAEB*, which carries the *EAEB* gene, was used to transform chemically competent *E. coli* strain BL-21. Because the influence of expression conditions on product distribution was well understood from the circularization study, we examined only three conditions: 2XYT medium at 16 °C with no induction (Condition I) and with 10 μM IPTG induction (Condition II), and at 37 °C with 1 mM IPTG induction (Condition III). The target protein was purified and characterized just as for AB20D. The results are summarized in Figure 3.

Figure 3a shows the SDS-PAGE analysis of the protein products obtained under the three conditions. In each of the three products (lanes 1−3), a prominent band appears near the 40 k MW marker. We assign this band to the domain-selective cyclized, tadpole protein. The MALDI-TOF mass spectrum of product EAEB-1 (Figure 3c) shows a MW of 32 161, close to that expected for the cyclized protein (32 154). There are also oligomers present in each lane as indicated by the arrows. Because the protein synthesis rate is slowest in Condition I and fastest in Condition III, it is reasonable to see increased amounts of chain-extended products in lanes 2 and 3. Using MALDI-TOF mass spectrometry, we identified dimeric and trimeric products in EAEB-3 (Figure S7). Notably, the chain-extended products in this case are either cyclic or linear proteins with multiple ELP arms—"comb-like" protein topologies that have not been reported previously.[35−37]

The cyclized topology of the major monomeric species is conveniently proven by proteolytic digestion at the TEV site. The digested product was analyzed by SDS-PAGE as shown in Figure 3a, lane 4. The digested product *l*-EAEB-1 exhibits reduced mobility in SDS-PAGE, a manifestation of the cyclic topology of its precursor EAEB-1. Moreover, the MALDI-TOF mass spectrum of the digested product (Figure 3d) remains essentially identical and matches that of EAEB-1 within the error of the MALDI-TOF mass spectrometry. In the SEC overlay (Figure 3b), the digested/relinearized sample *l*-EAEB-1 exhibits a symmetric elution profile at a retention volume much lower than EAEB-1. All of the evidence provides support for cyclization of only the domain flanked by SpyTag and SpyCatcher, as expected from the results obtained with AB20D.

The SEC profile shown in Figure 3b is significantly different from that of AB20D, in that EAEB exhibits a much stronger tendency toward cyclization. Under Condition I, EAEB-1 contains approximately 96% monomeric cyclized protein. With 10 μM IPTG induction (Condition II), 94% of the product EAEB-2 is monomeric cyclized protein. Even under Condition III, the mixture is ~82% monomeric tadpole protein, as estimated from integration of the area under the curve. Thus, the approach described here enables preparation of tadpole proteins in high purity, with minimum contamination by chain-extended products.

What accounts for the different cyclization efficiencies observed for AB20D and EAEB? We speculate that there are two main reasons. First, SpyTag is closer to SpyCatcher in EAEB than in AB20D, and the ring size is thus smaller. Jacobson−Stockmayer theory predicts that the cyclization probability diminishes with increasing ring size based on a random-flight model.[38] The model may not be directly applicable to cyclization by SpyTag−SpyCatcher chemistry since the preorganization of reactive groups via molecular recognition facilitates the cyclization process. Nevertheless, our results are qualitatively consistent with the prediction. Second, when the SpyTag is placed in the middle of the chain, it is buried within the random coil of the two tethered ELPs, which may hinder binding between SpyTag and the bulky SpyCatcher.

Topological Diversification *in Vitro*. The work described in the previous two sections focused on the preparation of unconventional protein topologies via *in situ* post-translational modification. It may be possible to extend this approach by expressing two or more mutually reactive telechelic proteins simultaneously in cells for reaction *in situ* to give complex topologies such as stars or branched proteins. However, because careful control of expression levels will be needed to ensure the desired reaction stoichiometry, we decided to prepare and purify the telechelic proteins separately for subsequent reaction in a first proof-of-concept. Scheme 3 shows several simple examples. The target topologies include a block protein formed by ligating terminally functionalized proteins (EA and EB), 3-arm star proteins made by linking a terminally functionalized protein and an internally functionalized protein (EA and EBE, or EB and EAE), a 4-arm star protein made by linking two internally functionalized proteins (EAE and EBE), and H-shaped proteins formed by conjugating a bifunctional protein with an internally monofunctionalized protein (AA and EBE, or BB and EAE). The folded CnaB2 domain that results from the reaction of SpyTag and SpyCatcher serves as the core of the star proteins and the branch junctions in the H-shaped proteins. The arm-number or H-shape thus refers to the ELP chains extending out from these cores or junctions.

The pQE plasmids carrying the telechelic protein genes of interest were used to transform *E. coli* strain BL-21. Proteins were expressed in 2XYT medium at 30 °C with 1 mM IPTG for 3−4 h. Proteins containing SpyTag were purified under denaturing conditions; those containing SpyCatcher were purified under native conditions. After dialysis and lyophilization, proteins were obtained as white powders with yields ranging from 25 to 40 mg/L. Purified telechelic proteins were analyzed by SDS-PAGE as shown in Figure 4. Mass spectrometry (Table 1, see also Figures S9, S11, and S13) shows that the MW of each protein is within 0.1% of the calculated value. These results confirm the purity and identity of each target telechelic protein.

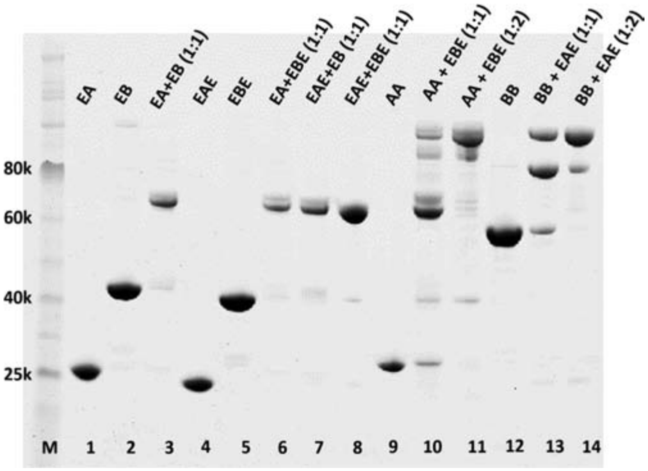


Figure 4. SDS-PAGE analysis of telechelic proteins and reaction products. Lane M is the MW marker. Lanes 1 and 2 are terminally functionalized proteins EA and EB. Lanes 4 and 5 are internally functionalized proteins EAE and EBE. Lanes 9 and 12 are bifunctional proteins AA and BB. Lane 3 is the block protein obtained by reacting EA and EB in a 1:1 ratio. Lanes 6 and 7 are the 3-arm star proteins obtained by reacting EA (or EB) and EBE (or EAE) in a 1:1 ratio. Lane 8 is the 4-arm star protein made by reacting EAE and EBE in a 1:1 ratio. Lanes 10 and 11 show the product distribution from the reaction between AA and EBE in 1:1 and 1:2 ratios, respectively. Lanes 13 and 14 show the product distribution from the reaction between BB and EAE in 1:1 and 1:2 ratios, respectively. Lanes 11 and 14 are predominantly H-shaped proteins.

In a first test case, we coupled terminally functionalized EA and EB to give a block protein conjugate that is analogous to a block copolymer[39−41] or a fusion protein. The synthesis of fusion proteins by conventional genetic engineering methods requires an N- to C-terminal (head-to-tail) junction, because the two coding sequences must be in the same reading frame. There is no such limitation in using SpyTag−SpyCatcher chemistry. One can link the proteins together in any way defined by the location of SpyTag and SpyCatcher. This means that not only is head-to-tail (from N- to C-terminus) conjugation possible, but head-to-head (from N- to N-terminus), tail-to-tail (from C- to C-terminus), and even body-to-body (from internal sites to internal sites) fusions are feasible. Here, the reaction between EA and EB represents a simple tail-to-tail conjugation. The remarkable efficiency of SpyTag−SpyCatcher chemistry leads to a clean and rapid ligation that is complete within 4 h. When the two components are mixed in equimolar amounts, the residual starting material is minimal in the final reaction mixture, and there is mainly only one reaction product, as shown by SDS-PAGE (Figure 4, lane 3). In the SEC overlay (Figure 5a), the product elutes at a much lower retention volume compared to the starting materials, consistent with the increased MW. The tailing on the low-MW side of the elution profile can be attributed to residual reactants. The reaction is nearly quantitative, as shown by the trace amounts of reactants remaining. The block protein was also characterized by MALDI-TOF mass spectrometry (Figure S14); the m/z value of 51 508 agrees with the expected value of 51 543 within the error of the measurement.

Table 1. Summary of the Telechelic Proteins and Proteins of Various Topologies

sample	construct	MW (calcd)	m/z (obs)	yield
l-AB20D	SpyTag−elastin−SpyCatcher	34002	34014	−
c-AB20D	SpyTag−elastin−SpyCatcher	33984	33973	~17 mg/L[a]
AB20A	SpyTag′−elastin−SpyCatcher	33958	33960	~20 mg/L[b]
l-EAEB	elastin−SpyTag− elastin−SpyCatcher	32172	32166	−
c-EAEB	elastin−SpyTag- elastin−SpyCatcher	32154	32161	~12 mg/L[a]
EA	elastin−SpyTag	19242	19259	~30 mg/L[c]
EB	elastin−SpyCatcher	32319	32324	~35 mg/L[b]
EAE	elastin−SpyTag−elastin	18409	18405	~25 mg/L[c]
EBE	elastin−SpyCatcher−elastin	31486	31463	~38 mg/L[b]
AA	SpyTag−elastin−SpyTag	20925	20926	~30 mg/L[c]
BB	SpyCatcher−elastin−SpyCatcher	47080	47065	~24 mg/L[b]
block ELP	EA+EB	51543	51508	~84%[d]
3-arm star ELP **1**	EA+EBE	50710	50731	~87%[d]
3-arm star ELP **2**	EAE+EB	50710	50719	~84%[d]
4-arm star ELP	EAE+EBE	49877	49870	~96%[d]
H-shaped ELP **3**	AA+2EBE	83862	83903	~85%[d]
H-shaped ELP **4**	BB+2EAE	83862	83865	~81%[d]

[a] Expression yield at 16 °C in 2XYT with 10 μM IPTG induction for 12 h. [b] Expression yield at 30 °C in 2XYT with 1 mM IPTG induction for 4 h. [c] Expression yield at 37 °C in 2XYT with 1 mM IPTG induction for 4 h. [d] Yields are based on densitometry analysis of SDS-PAGE gels.

While SpyTag is known to be reactive both terminally and internally, SpyCatcher has only been shown to react terminally.[21] We envisioned that as long as SpyCatcher is

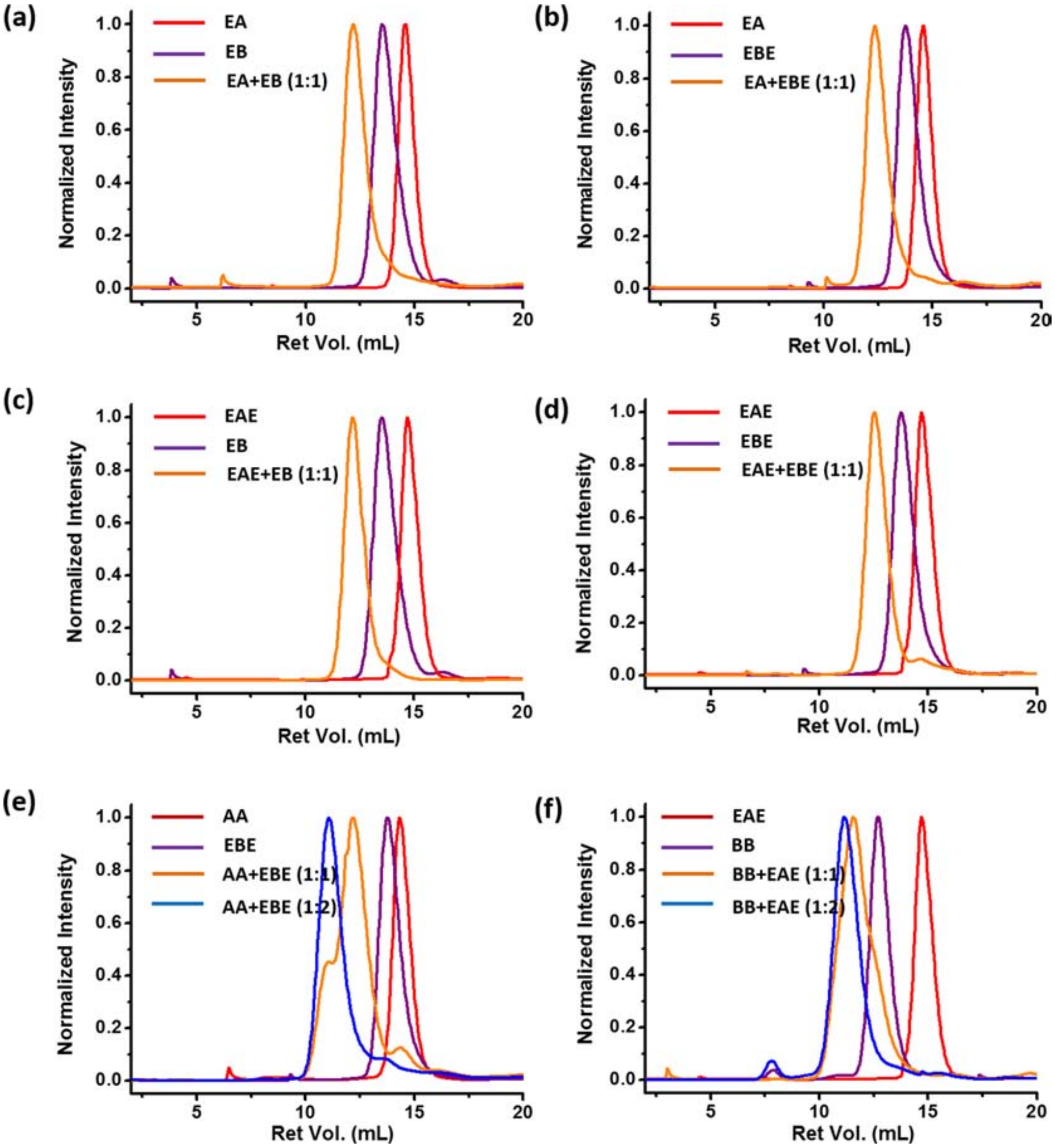


Figure 5. SEC overlay showing the progression of reaction: (a) EA+EB in 1:1 ratio; (b) EA+EBE in 1:1 ratio; (c) EAE+EB in 1:1 ratio; (d) EAE +EBE in 1:1 ratio; (e) AA+EBE in 1:1 and 1:2 ratios; and (f) BB+EAE in 1:1 and 1:2 ratios.

properly folded, the binding pocket should be available for SpyTag. Thus, SpyCatcher should also be reactive internally. Indeed, the reaction between terminally functionalized EA and internally functionalized EBE also proceeded efficiently (Figure 4, lane 6 in SDS-PAGE; Figure 5b, SEC overlay; and Figure S15a, MALDI-MS). The reaction between EAE and EB also proceeded cleanly as expected (Figure 4, lane 7 in SDS-PAGE; Figure 5c, SEC overlay; and Figure S15b, MALDI-MS). The products of these two reactions are a pair of 3-arm star protein isomers of identical MW (Table 1). A 4-arm star protein can also be conveniently prepared by mixing equimolar amounts of EAE and EBE. Even though both reactive units are internal, reactivity remains high. The reaction was complete within a few hours, and a single band appeared in SDS-PAGE with minimal residual EAE and EBE (Figure 4, lane 8). The shift in retention volume in the SEC overlay (Figure 5d) was also as expected. Confirmed by MALDI-MS (Table 1, Figure S15c), the product has a MW of 49 877. The block protein, 3-arm star proteins, and 4-arm star proteins are effectively topological isomers since their compositions and MW are nearly identical (ΔMW = 833).

More complex architectures are possible using multifunctional telechelic proteins. For example, H-shaped proteins can be obtained by reacting a bifunctional protein (AA or BB) and a complementary internally functionalized protein (EBE or EAE). We examined such reactions with 1:1 and 1:2 stoichiometries between the homotelechelic proteins. From both SDS-PAGE analysis and the SEC overlay, it is apparent that the equimolar reaction leads to a mixture of diadduct, monoadduct, and unreacted bifunctional protein, with the major product being the monoadduct. The identity of the monoadduct was confirmed by MALDI-MS of the reaction mixture (Figure S16c,d). When two equivalents of monofunctional protein were used (1:2 ratio), the reaction mixture contained predominantly the diadduct, or H-shaped protein. The H-shaped proteins **3** and **4** have exactly the same MWs (Table 1) and are also isomers. These results again confirm the high specificity and efficiency of the reaction.

Evidence of different topologies can be obtained by comparing SEC elution profiles of star- and H-shaped proteins (Figure 6a). The correlation between SEC elution volume and

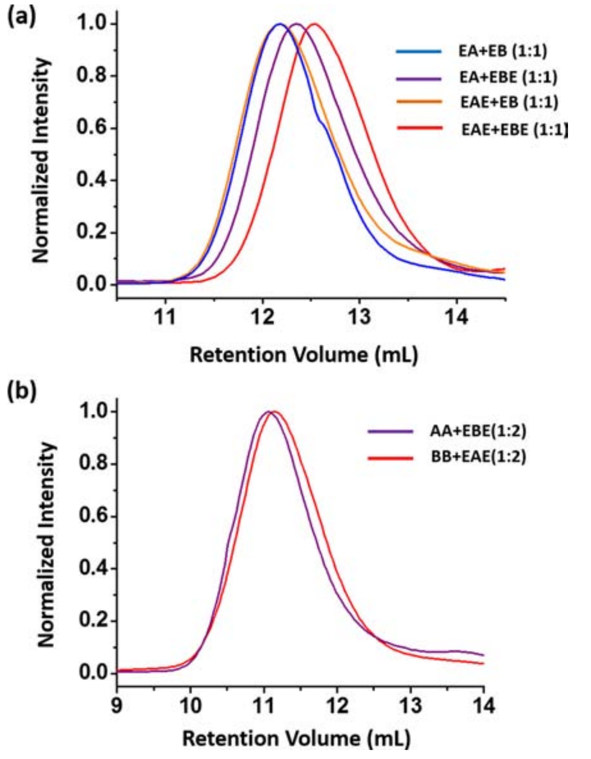


Figure 6. Comparison of the elution profiles of the topological protein isomers in SEC. (a) Elution profiles of block protein (blue), 3-arm star proteins **1** (purple) and **2** (brown), and 4-arm star protein (red). (b) Elution profiles of H-shaped proteins **3** (purple) and **4** (red).

the hydrodynamic volume of polymers is well established.[42] We anticipated that, despite the fact that the MW is held nearly constant, the hydrodynamic volume would decrease as the molecular topology changed from block to 3-arm star to 4-arm star, leading to corresponding increases in retention volume. This is generally true, as shown in Figure 6a. In addition, there are two observations that are somewhat counterintuitive. First, the elution profile of the block protein almost overlaps with that of the 3-arm star protein **2** (from EAE+EB), even though the block protein has a slightly higher MW and is expected to adopt a more expanded conformation. Second, the 3-arm star isomers, which have exactly the same MW, elute at distinct retention volumes, indicating that the isomers are characterized by different hydrodynamic volumes. In contrast, the SEC elution profiles of the H-shaped protein isomers almost overlap with one another. These observations can all be rationalized by the disposition of the ELP arms tethered to the folded CnaB2 domain formed by SpyTag and SpyCatcher (Figure 7).

The CnaB2 domain formed by SpyTag and SpyCatcher is a β-barrel structure, and the locations of the N- and C-termini for both SpyTag and SpyCatcher are fixed, as shown in the box in Figure 7. Both termini of SpyCatcher are located on the same end of the β-barrel. The N-terminus of SpyTag is located on the same end of the β-barrel as SpyCatcher, whereas its C-terminus is located on the opposite end. Due to this unique geometry, the tethered ELP arms may be congested on the same side of the β-barrel, leading to a more compact conformation than expected, such as that in block protein (EA+EB) and 3-arm star protein **1** (EA+EBE). Apparently, from Figure 7b, the 3-arm star protein **2** (EAE+EB) is more expanded than its isomer protein **1** (EA+EBE). The two H-shaped proteins seem to adopt similar configurations in solution.

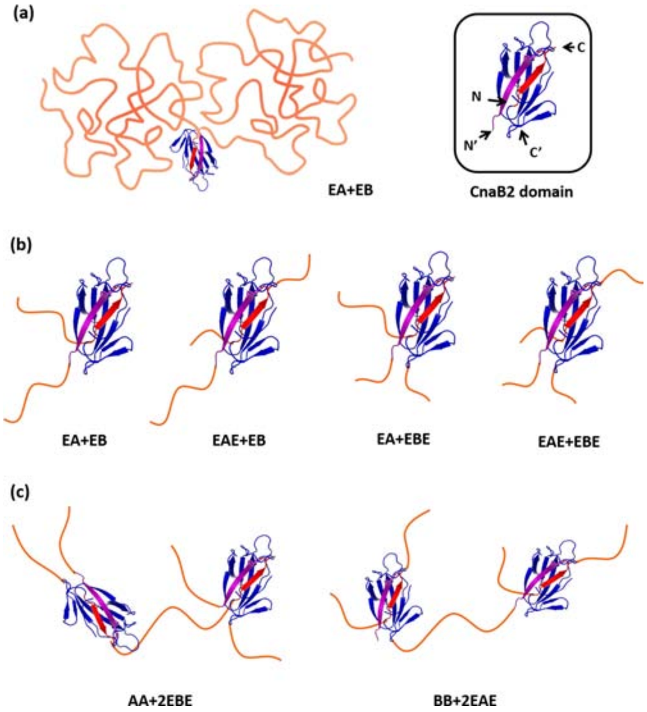


Figure 7. Configuration of ELP arms around the folded CnaB2 domain determines the hydrodynamic volume of branched proteins. The original CnaB2 domain is shown in the box. The SpyTag is shown in red, and the N- and C-termini are marked by N and C with arrows. The β-strand containing the Lys[31] in SpyCatcher is shown in purple, and the rest of SpyCatcher is shown in blue. The N- and C-termini of SpyCatcher are marked by N′ and C′ with arrows. The brown chain is ELP. The conformation of block protein (a) is drawn to scale. The ELP chains are 16 times longer than the length of the β-barrel of the folded CnaB2 domain. The cartoons of block protein and star proteins (b) and H-shaped proteins (c) are not drawn to scale in order to illustrate the difference in chain disposition on the β-barrel.

■ CONCLUSIONS

In summary, we demonstrate that genetically encoded SpyTag–SpyCatcher chemistry can be used to prepare unconventional protein topologies through either *in situ* post-translational modification or *in vitro* reaction. Circular, tadpole, block, star, and H-shaped proteins have been synthesized and characterized. Several related protein topologies, cyclic proteins with tethered multiple ELP arms and linear proteins with regular ELP side chains (protein "combs"), were also obtained as side products. The modular character of the SpyTag–SpyCatcher strategy should make it useful for preparing nonlinear macromolecules of diverse sequence and structure. There is no fundamental limitation on the MW of the proteins made by this strategy as long as the SpyCatcher domain remains soluble and folded. When combined with the machinery for cellular protein synthesis, SpyTag–SpyCatcher chemistry allows control of all of the fundamental properties of the macromolecular framework (length, sequence, stereochemistry, and topology), and provides a versatile new platform for the development of novel biomaterials.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Full amino acid sequences of the telechelic proteins, MALDI-TOF mass spectra of the telechelic proteins and the coupled proteins of various topologies, and the summary of bacterial strains and plasmids. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

tirrell@caltech.edu
frances@cheme.caltech.edu

Author Contributions

†W.-B.Z. and F.S. contributed equally to the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by National Science Foundation (DMR 1206121) and by the Gordon and Betty Moore Foundation through Grant GBMF2809 to the Caltech Programmable Molecular Technology Initiative.

■ REFERENCES

(1) Ober, C. K.; Cheng, S. Z. D.; Hammond, P. T.; Muthukumar, M.; Reichmanis, E.; Wooley, K. L.; Lodge, T. P. *Macromolecules* **2009**, *42*, 465−471.

(2) Hadjichristidis, N., Ed. *Complex Macromolecular Architectures: Synthesis, Characterization, and Self-assembly*; John Wiley & Sons: Hoboken, NJ, 2011.

(3) Tezuka, Y., Ed. *Topological Polymer Chemistry: Progress of Cyclic Polymers in Syntheses, Properties, and Functions*; World Scientific Publishing Co.: Hackensack, NJ, 2013.

(4) Trabi, M.; Craik, D. J. *Trends Biochem. Sci.* **2002**, *27*, 132−138.

(5) Conlan, B. F.; Gillon, A. D.; Craik, D. J.; Anderson, M. A. *Biopolymers* **2010**, *94*, 573−583.

(6) Craik, D. J. *Science* **2006**, *311*, 1563−1564.

(7) Wikoff, W. R.; Liljas, L.; Duda, R. L.; Tsuruta, H.; Hendrix, R. W.; Johnson, J. E. *Science* **2000**, *289*, 2129−2133.

(8) Kim, H. T.; Kim, K. P.; Lledias, F.; Kisselev, A. F.; Scaglione, K. M.; Skowyra, D.; Gygi, S. P.; Goldberg, A. L. *J. Biol. Chem.* **2007**, *282*, 17375−17386.

(9) Hochstrasser, M. *Nature* **2009**, *458*, 422−429.

(10) Kirisako, T.; Kamei, K.; Murata, S.; Kato, M.; Fukumoto, H.; Kanie, M.; Sano, S.; Tokunaga, F.; Tanaka, K.; Iwai, K. *EMBO J.* **2006**, *25*, 4877−4887.

(11) Camarero, J. A.; Pavel, J.; Muir, T. W. *Angew. Chem., Int. Ed.* **1998**, *37*, 347−349.

(12) Cowper, B.; Craik, D. J.; Macmillan, D. *ChemBioChem.* **2013**, *14*, 809−812.

(13) Jackson, D. Y.; Burnier, J. P.; Wells, J. A. *J. Am. Chem. Soc.* **1995**, *117*, 819−820.

(14) Daly, N. L.; Love, S.; Alewood, P. F.; Craik, D. J. *Biochemistry* **1999**, *38*, 10606−10614.

(15) Ingale, S.; Dawson, P. E. *Org. Lett.* **2011**, *13*, 2822−2825.

(16) Scott, C. P.; Abel-Santos, E.; Wall, M.; Wahnon, D. C.; Benkovic, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13638−13643.

(17) Iwai, H.; Lingel, A.; Pluckthun, A. *J. Biol. Chem.* **2001**, *276*, 16548−16554.

(18) Iwai, H.; Pluckthun, A. *FEBS Lett.* **1999**, *459*, 166−172.

(19) Antos, J. M.; Popp, M. W. L.; Ernst, R.; Chew, G. L.; Spooner, E.; Ploegh, H. L. *J. Biol. Chem.* **2009**, *284*, 16028−16036.

(20) Zakeri, B.; Howarth, M. *J. Am. Chem. Soc.* **2010**, *132*, 4526−4527.

(21) Zakeri, B.; Fierer, J. O.; Celik, E.; Chittock, E. C.; Schwarz-Linek, U.; Moy, V. T.; Howarth, M. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, E690−E697.

(22) Liu, J. C.; Heilshorn, S. C.; Tirrell, D. A. *Biomacromolecules* **2004**, *5*, 497−504.

(23) Urry, D. W. *J. Phys. Chem. B* **1997**, *101*, 11007−11028.

(24) Simnick, A. J.; Lim, D. W.; Chow, D.; Chilkoti, A. *Polym. Rev.* **2007**, *47*, 121−154.

(25) Laurent, B. A.; Grayson, S. M. *Chem. Soc. Rev.* **2009**, *38*, 2202−2213.

(26) Bielawski, C. W.; Benitez, D.; Grubbs, R. H. *Science* **2002**, *297*, 2041−2044.

(27) Oike, H.; Imaizumi, H.; Mouri, T.; Yoshioka, Y.; Uchibori, A.; Tezuka, Y. *J. Am. Chem. Soc.* **2000**, *122*, 9592−9599.

(28) Laurent, B. A.; Grayson, S. M. *J. Am. Chem. Soc.* **2006**, *128*, 4238−4239.

(29) Farewell, A.; Neidhardt, F. C. *J. Bacteriol.* **1998**, *180*, 4704−4710.

(30) Lemaux, P. G.; Herendeen, S. L.; Bloch, P. L.; Neidhardt, F. C. *Cell* **1978**, *13*, 427−434.

(31) Malakar, P.; Venkatesh, K. V. *Appl. Microbiol. Biotechnol.* **2012**, *93*, 2543−2549.

(32) Ali, M. H.; Imperiali, B. *Bioorg. Med. Chem.* **2005**, *13*, 5013−5020.

(33) Jia, Z. F.; Monteiro, M. J. *J. Polym. Sci. Part A: Polym. Chem.* **2012**, *50*, 2085−2097.

(34) Siebold, C.; Erni, B. *Biophys. Chem.* **2002**, *96*, 163−171.

(35) Advincula, R. C., Ed. *Polymer Brushes: Synthesis, Characterization, Applications*; Wiley-VCH: Weinheim, 2004.

(36) Johnson, J. A.; Lu, Y. Y.; Burts, A. O.; Xia, Y.; Durrell, A. C.; Tirrell, D. A.; Grubbs, R. H. *Macromolecules* **2010**, *43*, 10326−10335.

(37) Mittal, V., Ed. *Polymer Brushes: Substrates, Technologies, and Properties*; Taylor & Francis: Boca Raton, FL, 2012.

(38) Jacobson, H.; Stockmayer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600−1606.

(39) Hadjichristidis, N.; Pispas, S.; Floudas, G., Eds. *Block Copolymers: Synthetic Strategies, Physical Properties, and Applications*; Wiley-Interscience: Hoboken, NJ, 2003.

(40) Abetz, V., Ed. *Block Copolymers II*, Advances in Polymer Science 190; Springer: Berlin, 2005.

(41) Abetz, V., Ed. *Block Copolymers*, Advances in Polymer Science 189−190; Springer: Berlin, 2005.

(42) Grubisic, Z.; Rempp, P.; Benoit, H. *J. Polym. Sci., Polym. Lett.* **1967**, *5*, 753−759.