Published in final edited form as:
Nat Mater. 2016 April ; 15(4): 419–424. doi:10.1038/nmat4521.

Combinatorial codon scrambling enables scalable gene synthesis and amplification of repetitive proteins

Nicholas C Tang[1] and **Ashutosh Chilkoti**[1]

[1]Department of Biomedical Engineering, Duke University, Durham, North Carolina, 27708, USA

Introduction

Most genes are synthesized using seamless assembly methods that rely on polymerase chain reaction (PCR)[1_3]. However, PCR of genes encoding repetitive proteins either fail or generate nonspecific products. Motivated by the need to efficiently generate novel protein polymers through high-throughput gene synthesis, here we report a codon-scrambling algorithm that enables the PCR-based gene synthesis of repetitive proteins by exploiting the codon redundancy of amino acids and finding the least-repetitive synonymous gene sequence. We also show that the codon-scrambling problem is analogous to the well-known traveling salesman problem[4], and obtain an exact solution to it by using De Bruijn graphs[5] and a modern mixed integer linear program (MILP) solver. As experimental proof of the utility of this approach, we use it to optimize the synthetic genes for 19 repetitive proteins, and show that the gene fragments are amenable to PCR-based gene assembly and recombinant expression.

Main

Owing to the exceptional diversity of peptide properties in nature, repetitive proteins have been designed that exhibit unique structural and biological properties[6,7]. The majority of these artificial proteins are bioinspired from fibrous elastomeric animal proteins[8_15] ranging from highly elastic proteins —elastin, abductin, and resilin— to the highly tough silk fibroin. Other repetitive proteins have been designed with a broad array of applications in mind: peptide drugs[16], genome-editing tools[17], ligand scaffolds[18], purification tags[19,20], protein binders[21,22], and hydrogel-forming proteins that adopt β-sheet[23,24], coiled-coil[24], or random coil[25] structures.

The synthesis of genes encoding highly repetitive polypeptides is one of the unsolved problems in synthetic biology. Fast, scalable, high-throughput methods for the synthesis of

Correspondence should be addressed to A.C. (; Email: ashutosh.chilkoti@duke.edu).

Author Contributions: N.C.T. designed and performed experiments, developed algorithms, and prepared the manuscript. A.C. designed experiments and prepared the manuscript.

The authors declare no competing financial interests.

Code availability. All code is available at http://chilkotilab.pratt.duke.edu/codon-scrambler.

non-repetitive genes are available to molecular biologists which involve the precise hybridization of a large number of complementary oligonucleotides, followed by PCR amplification of the DNA product[1,26]. However, the highly repetitive DNA sequences that encode highly repetitive polypeptides are not amenable to efficient gene synthesis by current methods because the different gene fragments are all highly complementary to each other, so that precise assembly is not possible. Repetitive genes are also inaccessible to *in vitro* amplification and other manipulations such as sequencing, mutagenesis, cloning, and enzymatic error correction[27]. This is because, similar to gene synthesis, these manipulations include annealing steps in which single stranded DNA would anneal out of register with each other at multiple sites (Fig. 1a, b)[28].

Hence, specialized methods for gene assembly of repetitive proteins have been developed, including random oligomerization[16,29,30], and recursive ligation methods[8,11,25,31,32]. However, these methods also have one or more serious limitations: (1) they produce multimers of heterogeneous size, with a distribution that is dependent on DNA concentration and other reaction conditions, (2) require many laborious iterations of cloning in order to generate genes with a desired number of repeats, and (3) do not ensure directional insertion during cloning.

For these reasons, these specialized methods are far more tedious and require considerable optimization. They are also throughput-limited by the repeated and intensive steps of cloning and colony screening. The production scale of repetitive genes hence continues to fall behind that of non-repetitive genes because they cannot leverage the recent technological advances in next-generation, automated gene synthesis. In the past 15 years, these advances, driven by advancements in synthetic biology, have enabled the market cost of commercial gene synthesis to drop by over 100-fold[33]. Therefore, the development of PCR-based repetitive gene synthesis is critical to efficiently synthesize repetitive genes with high throughput[26].

In order to synthesize repetitive proteins via commercially compatible gene assembly approaches, we have developed a codon scrambling algorithm that identifies the least repetitive synonymous coding sequence for any protein sequence. Given the exponentially large number of codon variants, the search for the least repetitive sequence in an immense sequence space is a computational challenge. Although this discrete optimization problem belongs to the non-deterministic polynomial-time (NP)-hard complexity class, we found that developing a modestly sized problem formulation was crucial to solving it without resorting to metaheuristic algorithms, which are approximate and usually non-deterministic[34].

The objective of the codon scrambling problem is to minimize potential cross-hybridization events, or off-target interactions, of a repetitive coding sequence during polymerase reactions. In this problem, a polymerase reaction is conceptualized as a set of local cross-hybridization events between repeated subsequences. The tendency of duplex formation is estimated by the exponential-transformed duplexing energy ΔG given by the Boltzmann weight $\exp(-\Delta G/RT)$, where T is the system's absolute temperature and R is the gas constant. We can then quantify tradeoffs between sequences with a scalarized objective function, defined by the linear combination of all subsequence repeats weighted by their Boltzmann weights.

We utilize the *codon De Bruijn subgraph*, an adaptation of the De Bruijn subgraph, to sparsely represent all feasible nucleotide sequences that encode a protein of interest. A k-mer De Bruijn subgraph of a 64 codon alphabet Σ_{CODON} is a directed graph in which each vertex represents a substring consisting of k codons, or $3k$ nucleotides. Each arc is a substring consisting of $k + 1$ codons, and connects its prefix and suffix vertices; for example, arcs between vertices corresponding to dicodons can be represented by tricodons, as shown by the overlapping nucleotide sequences encoding the amino acids GVP below:

$$GGCGTA \xrightarrow{GGCGTA\boldsymbol{CCT}} GTA\boldsymbol{CCT}, \quad (1)$$

where $k = 2$. In this sense, a path traversal across multiple arcs represents a *sliding window*, where a window of $k + 1$ codons incrementally advances across a sequence. We can then construct a graph such that feasible paths code the protein of interest, provided that the graph is defined by the order and composition of the amino acid sequence (Fig. 1c). The codon De Bruijn subgraph is a relaxed version of the De Bruijn subgraph which allows for substrings that appear more than once if they code for different parts of the amino acid sequence. Arcs can easily be deleted from the subgraph if they contain forbidden sequences such as homopolymeric stretches, strong secondary structures, internal ribosome binding sites (iRBS)[35], or restriction enzyme recognition sites of interest.

In this graph, each arc is decomposed into a set of subsequences which is used to calculate the arc's (non-additive) contribution to the objective function. Therefore, the fixed-length path with the minimum objective defines the least repetitive codon variant. We have found that the discrete optimization on the set of all feasible paths is analogous to the classical traveling salesman problem. This is due to our utilization of *subtour elimination constraints*[4], used to eliminate disjoint cycles from the set of feasible solutions, as well as the requirement of simple, fixed-length paths. We provide a detailed description of the MILP formulation in the Supplementary Note, in which we utilize the graph's network topology to reduce the size and difficulty of the problem. In its implementation, we use MATLAB (Mathworks) code to convert repetitive protein sequences into MILP problem instances, which are then optimized using the IBM ILOG CPLEX Optimization Studio 12.5.1, a modern state-of-the-art MILP solver (Supplementary Data).

To test the algorithm, we selected a diverse set of 19 repetitive proteins (Table 1). These repetitive protein sequences can be completely optimized with CPLEX; computational experiments show that global optimality is reached within times ranging from 460 ms for wheat gliadin (SPR16)[32] to 2,020 s for elastin like polypeptide (ELP)[36]. The optimization objective eliminates repeats, while favoring shorter repeats with weak hybridization energies. In contrast, if two or more oligonucleotides are concatemerized[8], the resulting coding sequence is highly repetitive. This is apparent from the identity dot plot of ELP, where long intense lines off the main diagonal indicate long repeats (Fig. 1d). After optimization, these lines disappear or break into smaller lines with nucleotide lengths of ≤8 bp.

Computational experiments show how CPLEX timings scale with the number of repeat units of the ELP motif (Fig. 1e). As expected, computational run times have an exponential relationship with the increasing maximum repeated nucleotide length. This nucleotide length determines the De Bruijn window length *k*, which is exponentially proportional to the overall size of the graph and MILP problem (Supplementary Note). In contrast, the objective function increases continuously with the increasing number of motif repeats. For reasonable values of *k*, experimentally-useful numbers of repeats can be readily optimized (Table 1).

To illustrate the broad applicability of codon scrambling, we show that optimization enables gene synthesis of a diverse set of repetitive coding sequences. First, we selected 5 of the optimized genes because they span a range of objective values: GLP-1, BRT17, ELP[V-30], AB12, and 35-H-6. Using polymerase cycling assembly (PCA), we assembled their optimized sequences and their random-codon variants from their corresponding 80–90 nt oligonucleotide building blocks (Supplementary Table 2). For GLP-1, we additionally assembled a purely repetitive nucleotide sequence that would otherwise result from concatemerization of a gene that encodes a monomer of GLP-1. Microfluidic electropherograms of the resulting PCA products show many nonspecific bands for not only the purely repetitive codon variant, but also the random-codon variants (Fig. 2a). Due to the lack of distinct bands, we were unable to sequence or clone these genes. On the other hand, the optimized codon variants had distinct bands, and could be cloned into the pET-24a(+) expression vector and sequence verified (Supplementary Figs. 5–9, and Supplementary Table 3). Gen9 Inc. and Genscript Inc. successfully synthesized the other 14 optimized sequences (Fig. 2b). In contrast, non-optimized random-codon variants were far too repetitive to pass their sequence complexity filters.

Additionally, we show that the codon scrambling approach can also be applied to repetitive proteins with variable regions like transcription activator-like effectors (TALEs) and native semi-repetitive proteins like spider-silk, as their genes are also commonly rejected from gene synthesis vendors and require even more specialized methods of synthesis[14,15,17]. The optimized sequences for these genes, TN3[17] and ADF-1[14], were successfully synthesized and sequenced (Supplementary Figs. 10–13 and 16, and Supplementary Data).

A rank-order of objective values from Figure 2a suggests that a soft threshold between 10^6 and 10^7 exists between success and failure of PCR-based techniques. Longer sequences with objective values greater than 10^7 run the risk of PCA failure. This limitation can be addressed by a wide array of PCR-based manipulations of codon-scrambled genes that facilitate the assembly of longer multi-kilobase genes from shorter ones in a single cloning step. To illustrate this, we show that Type IIS restriction sites can be introduced into the 5' flanking sequence of PCR primers so that they are integrated into PCR products, which are then assembled together with the Golden Gate assembly method to create even longer repetitive genes. A Golden Gate reaction can assemble as many as 10 DNA fragments in a linear order where approximately 90% of recombinant clones contain the desired construct[3]. We used this strategy to create 150 repeats of the GVGVP amino acid motif (ELP[V-150]), which had a minimum possible objective value of 5.4 × 10^7, from optimized and PCA-assembled monomers consisting of 30 repeats (ELP[V-30]) (Supplementary Fig. 1a, b). After PCR and Golden Gate assembly, DNA restriction analysis confirmed that an estimated 93% of recombinant clones contained the desired construct (Supplementary Figs. 1c, 14, and 15). The benefits of this method over recursive ligation methods[37] become greater with longer target gene lengths (Supplementary Table 6). For example, the synthesis of genes with lengths of 5 kbp require only two cloning steps by codon scrambling, while recursive ligation requires six cloning steps.

We further demonstrate that the optimized genes, obtained from PCA or commercial synthesis, could be PCR-amplified with gene-specific primers for Gibson cloning into new vectors (Supplementary Table 7)[2]. Microfluidic electropherograms of the PCR products (Fig. 2c) show few to no nonspecific bands. We seamlessly cloned these PCR-amplified genes from their original pET-24a(+) vectors into linearized pMAL-c5X vectors containing a C-terminal His_6-tag and an N-terminal Maltose Binding Protein (MBP) tag, for enhanced soluble expression in *Escherichia coli*.

We finally show that manipulating codon usage with our method does not stifle heterologous expression of these repetitive coding sequences in *E. coli*, as there is a concern that synonymous codon usage is known to affect protein expression levels[38]. Despite the omission of codon bias rules during the codon scrambling of repetitive proteins, we found sufficient expression of 18 MBP-tagged genes for downstream characterization experiments (Fig. 3a). We further confirmed full-length products by the presence of C-terminal His_6-tags via western blot analysis (Supplementary Fig. 2) and direct DNA sequencing from *E. coli* cultures (Supplementary Fig. 16). We also directly compared the protein yields of ELP[V-30] with codons optimized for assembly (ELP[V-30] opt) with the same ELP with highly repetitive codons that were selected based on *E. coli* codon usage bias (ELP[V-30] orig) (Supplementary Fig. 4)[8,38]. Relative quantification by gel densitometry suggests that the expression of ELP[V-30] opt is at least 80% that of ELP[V-30] orig. These overall findings for the expression of codon-scrambled genes may presumably be due to balanced codon usage which do not include many offending codons overall, as well as the presence of a N-terminal sequence with low mRNA folding energy, which promotes efficient translation initiation[39]. Notably, however, the initial expression of AB12 only produced truncated proteins. We located a high percentage of ATG and Shine-Dalgarno (SD)-like motifs in the AB12 gene, which suggested that the poor expression of AB12 was caused by recurrent internal ribosome binding sites (iRBS) (Supplementary Fig. 3). Recoding the gene to eliminate the recurrent iRBS restored expression levels likely due to the avoidance of translational stalling (Supplementary Fig. 16 and Supplementary Data)[35]. Finally, as a test of protein functionality, ELP genes produced in this study were expressed and His_6-tag-purified, and exhibited the expected thermally responsive phase separation that is characteristic of this class of protein polymers (Fig. 3b, c)[19].

In conclusion, this study provides a robust and general solution to the long-standing problem of the scalable synthesis of highly repetitive coding sequences that are amenable to further oligomerization and manipulation. The open access availability of the gene design and optimization software that can be implemented on a personal computer will enable researchers to rapidly design genes for a wide range of repetitive proteins that are easily amenable to further manipulation owing to their optimized, minimally repetitive sequence — using powerful gene manipulation techniques such as Golden Gate assembly, Gibson assembly, and PCR amplification. It also provides the research community with a set of gene sequences and genes (deposited into Addgene) of a diverse range of repetitive proteins that are immediately available for protein expression. This design methodology serves to democratize repetitive protein synthesis, by providing easy access to the design and synthesis of new structurally and functionally diverse protein-based materials.

Methods

Gene design and assembly with PCA

Genes were designed so that the coding region was flanked by two BsaI recognition sites. BsaI generates 5’ overhangs for the purpose of directional scarless cutting in Golden Gate cloning. This entire sequence was additionally flanked by forward and reverse universal primer recognition sequences for the purpose of downstream PCR amplification. Overlapping oligonucleotides were designed with GeneDesign[40] with a target overlap-length of 40 bp, assembly length of 90 bp, and T_m of 70 °C (Supplementary Table 2). 1.25 pmol each of 8–12 oligonucleotides (Integrated DNA Technologies) of approximately 90 bp length were assembled into full-length genes in a 50 μL reaction containing 250 μM of each deoxynucleotide triphosphate (dNTP), 1X Herculase II PCR buffer, and 0.5 μL Herculase II Fusion polymerase (Agilent Technologies). The reaction was incubated at 98 °C for 15 min, followed by 15 cycles at 98 °C for 20 s, 57–67 °C for 20 s and 72 °C for 30 s, and a final step at 72 °C for 3 min (Supplementary Table 4). 1 μL of the full-length PCA product was then amplified by PCR using terminal universal primers (Supplementary Table 2). The 50 μL PCR reaction contained 250 μM of each dNTP, 1X Herculase II PCR buffer, 0.25 μM each of universal forward and reverse primers and 0.5 μL Herculase II Fusion polymerase. The reaction was incubated at 98 °C for 15 min, followed by 15 cycles at 98 °C for 20 s, 64 °C for 20 s and 72 °C for 30 s, and a final step at 72 °C for 3 min. The product was diluted 1:50 and then visualized with microfluidic electrophoresis using the Agilent high sensitivity DNA kit and the 2100 Bioanalyzer instrument (Agilent Technologies). BsaI recognition sites, 4 bp overhangs (CGGT/GGCT), and the His_6-tag coding region were added to a modified pET-24a(+) vector with primers, pET24a FWD/REV 1, using PCR (Supplementary Table 5). The construction of this modified vector from pET-24a(+) (Novagen) has been previously documented[8]. 2.5 ng of the plasmid was linearized and amplified in a 50 μL PCR reaction with the same components as previously described. The reaction was incubated at 95 °C for 2 min, followed by 5 cycles at 95 °C for 20 s, 56 °C for 20 s and 72 °C for 2 min 40 s, followed by 25 cycles at 95 °C for 20 s and 72 °C for 3 min, and a final step at 72 °C for 3 min. The gene assembly product and linearized vector were gel-purified from a 1% and 0.4% low melting point agarose gel respectively using the QIAquick gel extraction kit (Qiagen). The concentration of the linearized vector was adjusted to 100 ng μL^{-1}, and gene assembly products were adjusted to a 1:1 molar ratio to the vector. 1 μL of each purified product was combined in a 10 μL Golden Gate digestion-ligation reaction containing 0.75 μL BsaI (20 U μL^{-1}; New England Biolabs), 1X NEBuffer 4, 1X bovine serum albumin (BSA), 0.25 μL T7 Ligase (3000 U μL^{-1}; Enzymatics), and 1 mM adenosine triphosphate

(ATP). The reaction was incubated for 20 cycles at 37 °C for 5 min and 20 °C for 5 min. The 7 μL of the Golden Gate product was treated with exonuclease to remove incomplete ligation products in 10 μL reaction consisting of 1 μL PlasmidSafe DNAse (10 U μL^{-1}; Epicentre), 1X PlasmidSafe Reaction Buffer, and 1 mM ATP. The reaction was incubated at 37 °C for 30 min, followed by inactivation at 70 °C for 30 min. 5 μL of the PlasmidSafe reaction product was transformed into 50 μL EB5α competent cells (EdgeBio) following the manufacturer's instructions. Successful clones were screened with direct sequencing from colonies using standard T7 primers. Plasmids were prepared from sequence-verified colonies using the QIAprep spin miniprep kit (Qiagen).

PCR and Golden Gate assembly of ELP

BsaI recognition sites, 4 bp overhangs (GGAG/ACCC), and the His_6-tag coding region were added to the modified pET-24a(+) vector with primers, pET24a FWD/REV 2, using PCR (Supplementary Table 5). ELP[V-30] was then gene-assembled with PCA and cloned into the modified pET-24a(+) vector as described in the previous section. 6 pairs of custom primers, 1–5 FWD/REV and pET24a FWD/REV 3, were designed to attach BsaI recognition sites with unique 4 bp overhangs so that all 5 PCR products and a linearized modified pET-24a(+) vector would ligate together in a single Golden Gate reaction to form the gene coding ELP[V-150] (Supplementary Table 5). Each 4 bp overhang was designed to be unique using codon degeneracy and/or 2 bp shifts in either direction (Supplementary Fig. 1a). 2.5 ng of ELP[V-150]-pET24a plasmid was amplified in 50 μL PCR reactions with the same components as previously described. The reactions were incubated at 98 °C for 3 min, followed by 30 cycles of 98 °C for 20 s, 61 °C for 20 s and 72 °C for 10 s, and a final step at 72 °C for 3 min. The PCR products were visualized and gel-purified from a 2% low melting point agarose gel using the QIAquick gel extraction kit. The concentration of each insert was estimated from the gel using the Image Lab software (BioRad). The inserts were each adjusted to 2.68 ng μL^{-1}, or 1:1 molar ratio with 100 ng μL^{-1} linearized vector. 1 μL of each of the 6 gel-purified PCR products were then assembled in a 10 μL Golden Gate digestion-ligation reaction, treated with PlasmidSafe exonuclease, and transformed into EB5α competent cells with components and incubation times as described previously. Plasmids were prepared from 14 colonies, and were subjected to DNA restriction analysis with XbaI and BamHI, followed by DNA sequencing.

ELP protein expression and purification

Modified pET-24a(+) plasmids containing ELP[V-30], ELP[V-60], ELP[V-150] and ELP[AV-60] were each transformed into BL21 Star competent cells (Life Technologies) following the manufacturer's instructions. Colonies were inoculated in 2 mL of Terrific Broth (TB) plus 100 μg mL^{-1} kanamycin in 14 mL round-bottom tubes and grown for 8 hours at 37 °C and 200 rpm. 100 μL of the starter cultures were inoculated in 10 mL of Overnight Express TB (EMD Millipore) plus 100 μg mL^{-1} kanamycin in 50 mL Bio-Reaction tubes (Celltreat) and grown for 24 h at 30 °C and 140 rpm. The cells were pelleted and lysed with 0.5 mL B-PER complete bacterial protein extraction reagent (Life Technologies) according to the manufacturer's instructions. 1X Halt protease inhibiter cocktail (Life Technologies) and 1 mM EDTA was added and the lysates were subject to 4 cycles of freeze-thaw. Lysates were purified using the HisExpress columns (ClaremontBio) following the manufacturer's instructions. 10 mL of 1X phosphate buffered saline (PBS) plus 10 mM imidazole was used as the wash buffer and 1 mL of 1X PBS plus 250 mM imidazole was used as the elution buffer. Protein yields were quantified with the Quant-iT protein assay kit (Life Technologies) and were diluted with 1X PBS to 20 μM concentrations. 150 pmol of each ELP was also separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) on a Mini-PROTEAN TGX stain-free precast gel (BioRad). The gels were exposed to 1 minute of 302 nm UV light to produce fluorescence and imaged with a Gel Doc XR system (BioRad). To characterize the inverse transition temperature of ELPs, the 350 nm optical densities of 25 μM ELP solutions were monitored with a Cary 300 UV-Vis spectrophotometer (Agilent Technologies) as a function of solution temperature.

PCR and Gibson assembly

Gene-specific Gibson assembly primers were designed to amplify the coding sequence (CDS) of each gene and add overhangs containing the vector-overlapping sequences, 5’-GAAGGTCTTCCGGT-3’ and 5’-GGCTCGGGACATCATC-3’ (Supplementary Table 7). 0.25 ng of pET24a(+) plasmids containing each gene were amplified with their respective primers in 50 μL PCR reactions with the same components as previously described. The reactions were incubated at 98 °C for 3 min, followed by 30 cycles of 98 °C for 20 s, 47.7–64.4 °C for 20 s and 72 °C for 30–60 s, and a final step at 72 °C for 3 min (Supplementary Table 4). The products were diluted 1:50 and then visualized with microfluidic electrophoresis using the Agilent high sensitivity DNA kit and the 2100 Bioanalyzer instrument. The pMAL-c5X vector (New England Biolabs) was linearized to contain the overlap sequences at each end with primers, pMAL-c5X FWD/REV, using PCR (Supplementary Table 7). All PCR products were then purified with the AxyPrep Mag PCR clean-up kit (Axygen) and quantified on a NanoDrop 1000 spectrophotometer (Thermo Scientific) by the absorbance at 260 nm. 30 ng of linearized pMAL-c5X and 60 fmol of the insert, at 3.5:1 molar ratio with the vector, were combined in a 10 μL Gibson assembly reaction containing 1X Gibson assembly master mix (New England Biolabs). The reaction was incubated at 40 °C for 1 h. 2 μL of a 1:4 dilution of each Gibson product were transformed into 50 μL EB5α competent cells following the manufacturer's instructions. Successful clones were screened with direct sequencing from colonies using pMAL sequencing primers (Supplementary Table 7). Plasmids were prepared from sequence-verified colonies using the QIAprep spin miniprep kit. All plasmids are available from Addgene (Plasmids #66996–#67015).

Recombinant protein expression

pMAL-c5X plasmids containing each gene were transformed into NEB Express competent cells (New England Biolabs) following the manufacturer's instructions. Colonies were inoculated and grown in 2 mL of TB plus 50 μg mL^{-1} carbenicillin, then inoculated and grown in 10 mL of Overnight Express TB plus 50 μg mL^{-1} carbenicillin, and finally pelleted and lysed as described previously for ELP protein expression. Lysates were diluted based on dry pellet weight to 6.67 μg μL^{-1} with 1X SDS plus 1X Halt protease inhibiter cocktail and 1 mM EDTA. 50 μg of each lysate was heated to 98 °C for 10 min and then separated by SDS-PAGE on Mini-PROTEAN TGX stain-free precast gels. The gels were exposed to 1 minute of 302 nm UV light to produce fluorescence and imaged with a Gel Doc XR system.

The presence of the C-terminal His$_6$-tag was verified with Western blot analysis, using a 1:4000 dilution of DyLight 650 conjugated 6x-His epitope tag antibody (Life Technologies, cat. no. MA1-21315-D650). The resulting western blot was imaged using a Typhoon 9410 (GE Healthcare).

Computation

The minimum hybridization energies for all substrings of each repetitive protein sequence were calculated with UNAFOLD 3.8 at a temperature of 50 °C and a Na^+ concentration of 50 mM. The sequences were converted into MILP problem instances with MATLAB R2012a (Mathworks) (Supplementary Data). These instances were then solved using a MATLAB interface to CPLEX 12.5.1 (IBM) using a Dell Precision M4700 Laptop with 8 GB RAM and a quad-core i7-3820QM 2.70 GHz processor (Supplementary Table 8). iRBS were found by calculating the minimum hybridization energies of all arc-sequences with the *E. coli* anti-SD sequence, 5'-ACCTCCTTA-3'. ΔG values were calculated with UNAFOLD 3.8 at a temperature of 37 °C and a Na^+ concentration of 50 mM. For AB12, arcs with values lower than -8 kcal mol^{-1} were deleted from the graph. Other forbidden sequences included BsaI, NdeI, BseRI, BamHI, and XbaI recognition sites, as well as stretches of 6 Gs, 8 Cs, 8 As, and 8 Ts (Supplementary Data).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank S. Mukherjee for valuable advice on mathematical optimization, and K. Dooley for useful discussions on soluble protein expression. This work was funded by the NIH through grant # GM061232 to A.C. and by the NSF through the Research Triangle MRSEC (NSF DMR-11-21107). N.C.T. was supported by a fellowship from the Duke University Center of Biomolecular and Tissue Engineering.

References

1. Ma S, Tang N, Tian J. DNA synthesis, assembly and applications in synthetic biology. Curr. Opin. Chem. Biol. 2012; 16:260–7. [PubMed: 22633067]
2. Gibson DG, et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods. 2009; 6:343–345. [PubMed: 19363495]
3. Engler C, Gruetzner R, Kandzia R, Marillonnet S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One. 2009; 4:e5553. [PubMed: 19436741]
4. Laporte G. The traveling salesman problem: An overview of exact and approximate algorithms. Eur. J. Oper. Res. 1992; 59:231–247.
5. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U. S. A. 2001; 98:9748–53. [PubMed: 11504945]
6. Kaplan, DL.; Mello, SM.; Arcidiacono, S.; Fossey, S,SK. Protein based materials. McGrath, K.; Kaplan, D., editors. Birkhäuser; Boston: 1998. p. 103-131.doi:10.1007/978-1-4612-4094-5
7. Cranford, SW.; Buehler, MJ. Biomateriomics. Vol. 165. Springer; 2012.
8. McDaniel JR, Mackay JA, Quiroz FG, Chilkoti A. Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes. Biomacromolecules. 2010; 11:944–52. [PubMed: 20184309]

9. Anderson, D.; Maugh, K. Escherichia coli expression vector encoding bioadhesive precursor protein analogs comprising three to twenty repeats of the decapeptide (Ala-Lys-Pro-Ser-Tyr-Pro. 1992. US Pat. 5,149,657

10. Lyons RE, et al. Design and facile production of recombinant resilin-like polypeptides: gene construction and a rapid protein purification method. Protein Eng. Des. Sel. 2007; 20:25–32. [PubMed: 17218334]

11. Su RS-C, Renner JN, Liu JC. Synthesis and characterization of recombinant abductin-based proteins. Biomacromolecules. 2013; 14:4301–8. [PubMed: 24147646]

12. Cappello, J.; Ferrari, F.; Richardson, C. Methods for preparing synthetic repetitive DNA.. 1997. US Pat. 5,641,648

13. Cappello, J.; Causey, S. Peptides comprising repetitive units of amino acids and DNA sequences encoding the same.. 2000. US Pat. 6,018,030 262

14. Widmaier DM, et al. Engineering the Salmonella type III secretion system to export spider silk monomers. Mol. Syst. Biol. 2009; 5:309. [PubMed: 19756048]

15. Tokareva O, Michalczechen-Lacerda VA, Rech EL, Kaplan DL. Recombinant DNA production of spider silk proteins. Microb. Biotechnol. 2013; 6:651–63. [PubMed: 24119078]

16. Amiram M, Quiroz F, Callahan D, Chilkoti A. A highly parallel method for synthesizing DNA repeats enables the discovery of 'smart' protein polymers. Nat. Mater. 2011; 10:141–8. [PubMed: 21258353]

17. Ousterout DG, et al. Reading frame correction by targeted genome editing restores dystrophin expression in cells from Duchenne muscular dystrophy patients. Mol. Ther. 2013; 21:1718–26. [PubMed: 23732986]

18. Farmer RS, Top A, Argust LM, Liu S, Kiick KL. Evaluation of conformation and association behavior of multivalent alanine-rich polypeptides. Pharm. Res. 2008; 25:700–8. [PubMed: 17674161]

19. McDaniel JR, Radford DC, Chilkoti A. A unified model for de novo design of elastin-like polypeptides with tunable inverse transition temperatures. Biomacromolecules. 2013; 14:2866–72. [PubMed: 23808597]

20. Shur O, Banta S. Rearranging and concatenating a native RTX domain to understand sequence modularity. Protein Eng. Des. Sel. 2013; 26:171–80. [PubMed: 23173179]

21. Steiner D, Forrer P, Plückthun A. Efficient selection of DARPins with subnanomolar affinities using SRP phage display. J. Mol. Biol. 2008; 382:1211–27. [PubMed: 18706916]

22. Lee BW, et al. Strongly binding cell-adhesive polypeptides of programmable valencies. Angew. Chemie - Int. Ed. 2010; 49:1971–1975.

23. Higashiya S, Topilina N. Design and preparation of β-sheet forming repetitive and block-copolymerized polypeptides. Biomacromolecules. 2007

24. Petka W, Harden J, McGrath K, Wirtz D, Tirrell D. Reversible hydrogels from self-assembling artificial proteins. Science (80−.). 1998; 281:389–392.

25. Davis NE, Ding S, Forster RE, Pinkas DM, Barron AE. Modular enzymatically crosslinked protein polymer hydrogels for in situ gelation. Biomaterials. 2010; 31:7288–97. [PubMed: 20609472]

26. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods. 2014; 11:499–507. [PubMed: 24781323]

27. Ma S, Saaem I, Tian J. Error correction in gene synthesis technology. Trends Biotechnol. 2012; 30:147–54. [PubMed: 22209624]

28. Hommelsheim CM, Frantzeskakis L, Huang M, Ülker B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci. Rep. 2014; 4:5052. [PubMed: 24852006]

29. O'Brien JP, et al. in Silk Polymers: Materials Science and Biotechnology. 1994; 544:104–117.

30. Kurihara H, Morita T, Shinkai M, Nagamune T. Recombinant extracellular matrix-like proteins with repetitive elastin or collagen-like functional motifs. Biotechnol. Lett. 2005; 27:665–70. [PubMed: 15977075]

31. Goeden-Wood NL, Conticello VP, Muller SJ, Keasling JD. Improved Assembly of Multimeric Genes for the Biosynthetic Production of Protein Polymers. Biomacromolecules. 2002; 3:874–879. [PubMed: 12099837]

32. Elmorjani K, et al. Synthetic genes specifying periodic polymers modelled on the repetitive domain of wheat gliadins: conception and expression. Biochem. Biophys. Res. Commun. 1997; 239:240–246. [PubMed: 9345302]

33. Carlson, R. Time for New DNA Synthesis and Sequencing Cost Curves. 2014. at <http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html>

34. Gendreau M, Potvin J-Y. Handbook of Metaheuristics. Handbook of Metaheuristics. 2010; 146

35. Whitaker WR, Lee H, Arkin AP, Dueber JE. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. ACS Synth. Biol. 2015; 4:249–57. [PubMed: 24931615]

36. Meyer DE, Trabbic-Carlson K, Chilkoti A. Protein purification by fusion with an environmentally responsive elastin-like polypeptide: Effect of polypeptide length on the purification of thioredoxin. Biotechnol. Prog. 2001; 17:720–728. [PubMed: 11485434]

37. Meyer DE, Chilkoti A. Genetically encoded synthesis of protein-based polymers with precisely specified molecular weight and sequence by recursive directional ligation: examples from the elastin-like polypeptide system. Biomacromolecules. 2002; 3:357–67. [PubMed: 11888323]

38. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. U. S. A. 2010; 107:3645–50. [PubMed: 20133581]

39. Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science. 2013; 342:475–9. [PubMed: 24072823]

40. Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD. GeneDesign: rapid, automated design of multikilobase synthetic genes. Genome Res. 2006; 16:550–6. [PubMed: 16481661]

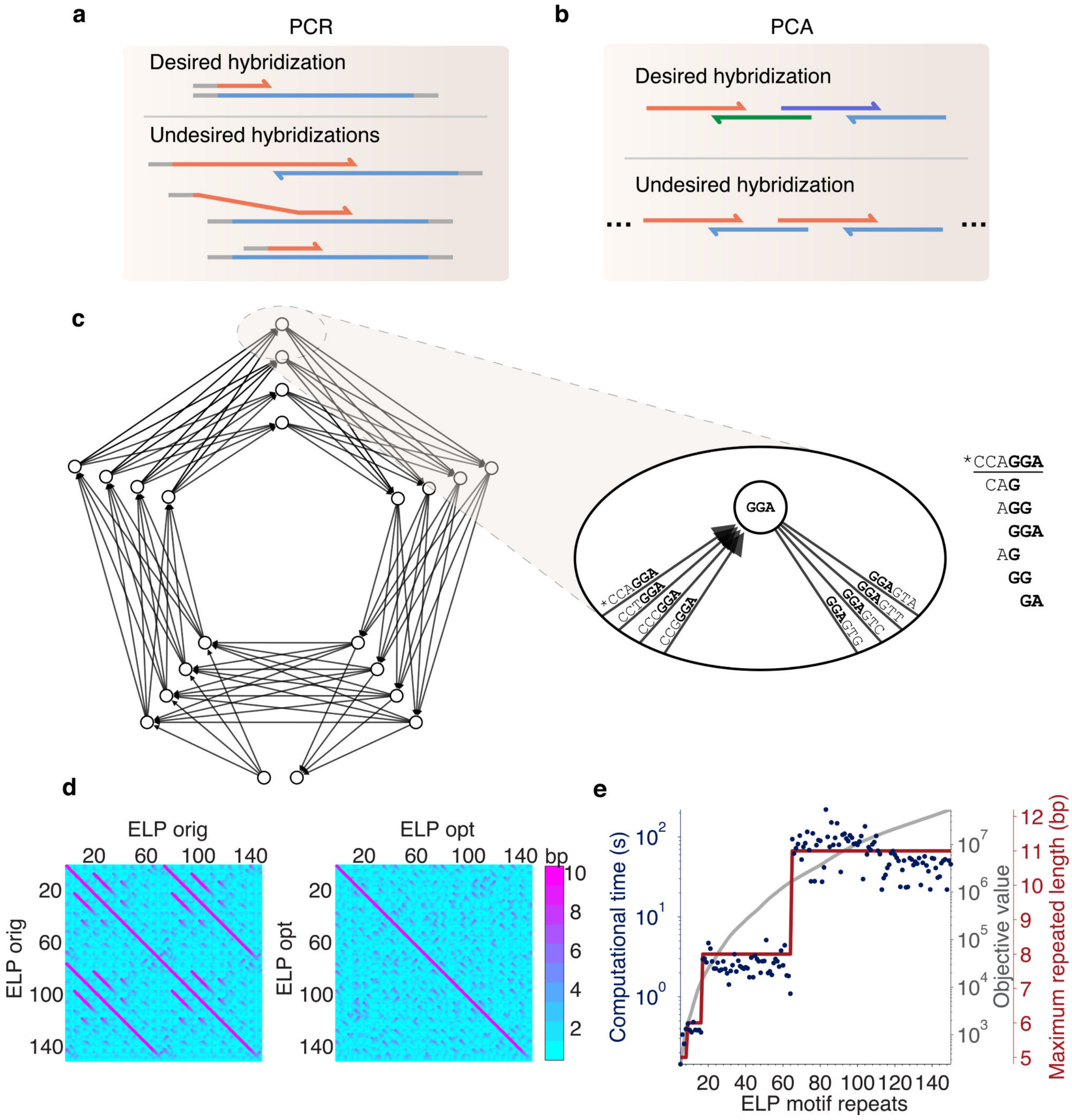

Figure 1. Computational analysis of codon scrambling

a, Schematic of possible hybridizations during the annealing step of PCR. The 3’ end of complete or incompletely extended single stranded fragments can hybridize out of register with each other, producing longer products. Also, primers can hybridize to non-specific sites, producing shorter products. **b**, Schematic of possible hybridizations during the annealing step of polymerase cycling assembly (PCA). End-complementarity between different pairs of oligonucleotide building blocks produces out of order assembly products. **c**, Example diagram of a cyclic codon De Bruijn subgraph with a window length *k* = 1 and a motif length $r = 5$ (GVGVP amino acid motif). Each arc is decomposed into subsequences which are used to calculate the objective function. **d**, Nucleotide sequence dot plot for 10 repeats of the GVGVP motif in ELP compared against itself, with codons that would result from concatemerization with two oligonucleotide inserts (ELP orig, left) and with optimal non-repetitive codons (ELP opt, right). Points describe exact sequence identity, including the reverse complement. **e**, Timings of CPLEX optimization for ELP with up to 150 motif repeats, with corresponding objective values and maximum repeated nucleotide lengths.

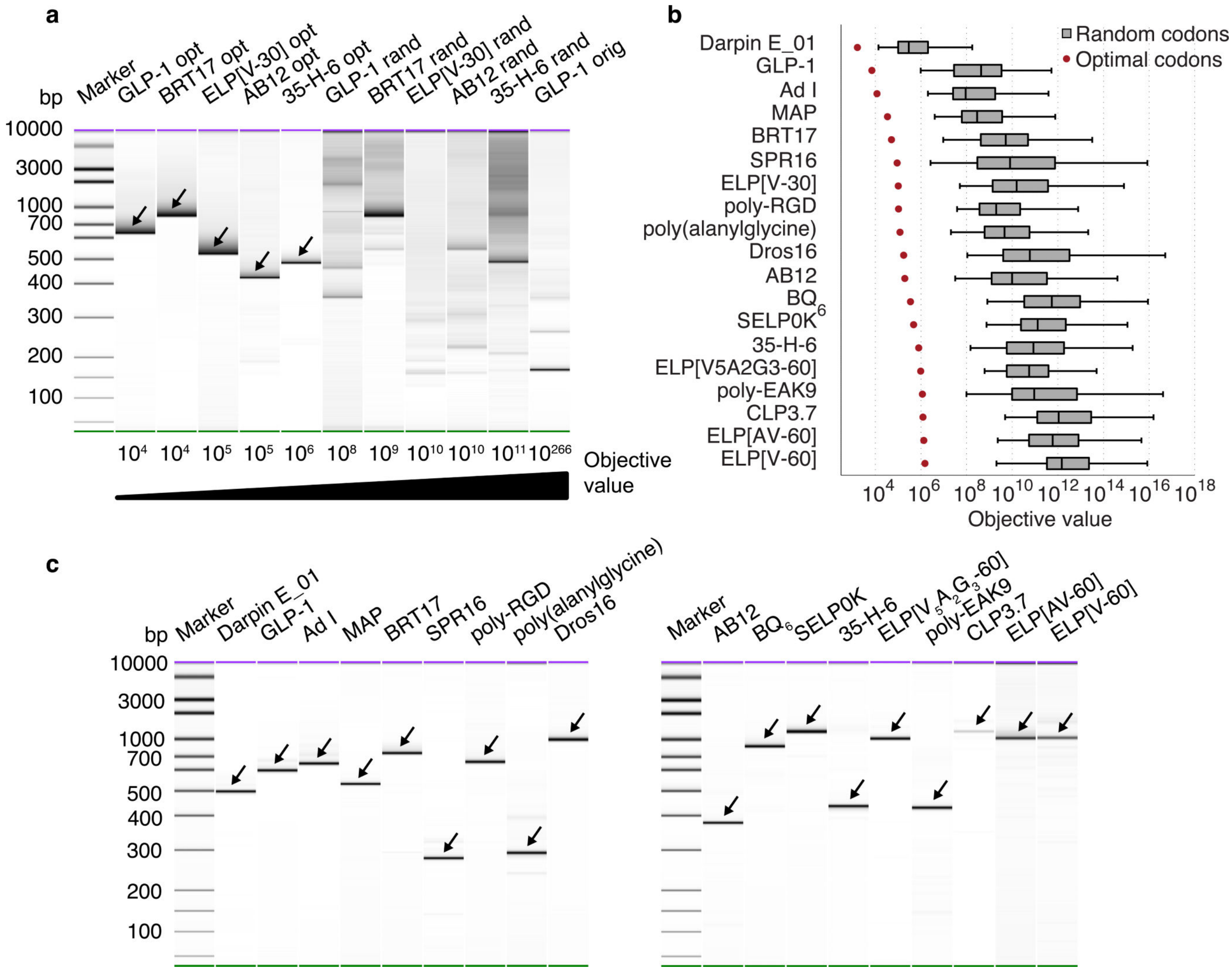

Figure 2. Gene assembly of a diverse set of repetitive proteins

a, Microfluidic electropherograms of PCA products, rank-ordered by calculated objective values. We assembled oligonucleotides into full-length genes with optimal codons (opt), randomized codons (rand) and original codons that would otherwise result from concatemerization (orig). The distinct bands marked by arrows represent assembled genes of correct length and sequence that are subsequently cloned into the pET-24a(+) vector. **b**, Rank-ordered objective values of the optimal-codon variant and 100 random-codon variants of each repetitive gene. Boxes signify the interquartile ranges (IQR) between the 25th and 75th percentiles and the lines signify the medians. **c**, Microfluidic electropherograms of PCR products of optimized genes. Each PCR reaction successfully amplified the gene with gene-specific primers for subsequent Gibson cloning. The distinct bands marked by arrows represent PCR products of correct lengths and sequences.

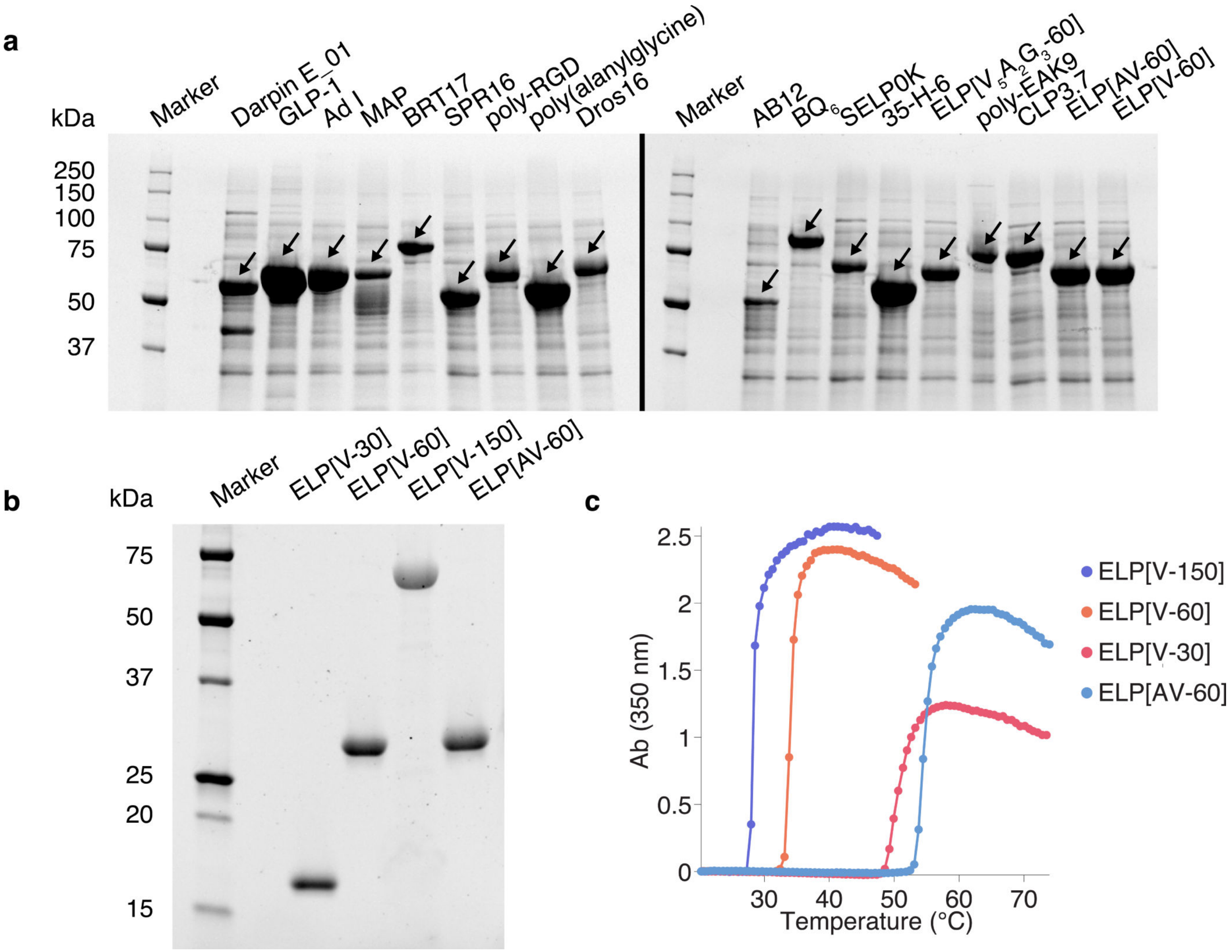

Figure 3. Protein expression of a diverse set of repetitive proteins

a, Recombinant expression of MBP-tagged repetitive proteins in *E. coli* from optimized gene sequences. 50 μg of total cell lysate for each protein were separated and visualized on tris-glycine extended (TGX) stain-free gels. The broad bands marked by arrows represent expressed repetitive proteins. Note that while CLP3.7 produced a relatively weak PCR band (Fig. 2c), the protein nevertheless expressed well. **b**, Purification of ELPs. 150 pmol of each purified ELP were separated on TGX stain-free gels. **c**, Turbidity profiles for purified ELPs with inverse phase transition behavior. Observed transition temperatures occur at 28.6 °C, 35.2 °C, 50.6 °C, and 55.2 °C for ELP[V-150], ELP[V-60], ELP[V-30], and ELP[AV-60] respectively. These transition temperatures are within 1 °C of model-predicted temperatures: 28.6 °C, 34.8 °C, 50.4 °C, and 55.8 °C[19].

Table 1

Computational results for the optimization of repetitive proteins

	Sequence	Objective value	Time (s)
Designed ankyrin repeat protein (Darpin E_01) [21]	DLGKKLLEAARAGQDDEVRILMANGADVNA DDTWGWTPLHLAAYQGHLEIVEVLLKNGADVNA YDYIGWTPLHLAADGHLEIVEVLLKNGADVNA)YIGDTP LHLAAHNGHLEIVEVLLKHGADVNA QDKFGKTAFDISIDNGNEDLAEILQ	1.65×10^3	684
Glucagon-like peptide (GLP-1) [16]	[GAHGEGTFTSDVSSYLEEQAAKEFIAWLVKGR]$_{6}$	7.09×10^3	10.9
Adenovirus-like construct (Ad I) [29]	[LSVQTSAPLTVSDGK]$_{14}$	1.20×10^4	72.2
Mussel adhesive protein (MAP) [9]	[AKPSYPPTYK]$_{16}$	3.41×10^4	1.33
Repeats in toxin (BRT17) [20]	[GGAGNDTLY]$_{17}$	5.04×10^4	31.3
Wheat gliadin (SPR16) [32]	[PQQPY]$_{16}$	9.23×10^4	0.460
Cell adhesive substrate (poly-RGD) [22]	[GSGSGSGRGDS]$_{20}$	1.05×10^5	79.5
β-sheet forming polypeptide (poly(alanylglycine)) [24]	[AGAGAGPEG]$_{10}$	1.22×10^5	5.06
Resilin like polypeptide (Dros 16) [10]	[GAPGGGNGGRPSDTY]$_{16}$	1.74×10^5	603
Abductin (AB12) [11]	[FGGMGGGNAGFGGMGGGKAGFGGMGGGNAG]$_{4}$	2.00×10^5	24.6
Transglutaminase substrate peptide (BQ$_6$) [25]	[[GQQQLGGAGTGSA]$_{2}$ [GAGQGEA]$_{3}$]$_{6}$	3.44×10^5	285
Silk-elastin like polypeptide (SELP0K) [12]	[[GAGAGS]$_{2}$ [GVGVP]$_{4}$GKGVP [GVGVP]$_{3}$]$_{6}$ [GAGAGS]$_{2}$	4.72×10^5	913
Alanine-rich polypeptides (35-H-6) [18]	[AAAQAAQAQAAAEAAAQAAQAQ]$_{6}$	8.11×10^5	67.35
Elastin like polypeptide (ELP[$V_5A_2G_3$-60]) [36]	[[GVGVP]$_{2}$GGGVPGAGVP [GVGVP]$_{3}$GGGVPGAGVPGGGVP]$_{6}$	9.77×10^5	2,020
β-sheet forming polypeptide (poly-EAK9) [31]	[AEAEAKAK]$_{18}$	1.16×10^6	2.45
Collagen like protein (CLP3.7) [13]	[GAPGTPGPQGLPGSP]$_{24}$	1.22×10^6	18.0
Elastin like polypeptide (ELP[AV-60]) [19]	[GAGVP GVGVP]$_{30}$	1.35×10^6	14.1
Elastin like polypeptide (ELP[V-60]) [36]	[GVGVP]$_{60}$	1.48×10^6	4.02