SYNTHETIC BIOLOGY

Synthetic sequence entanglement augments stability and containment of genetic information in cells

Tomasz Blazejewski[1,2]*, Hsing-I Ho[1]*, Harris H. Wang[1,3]†

In synthetic biology, methods for stabilizing genetically engineered functions and confining recombinant DNA to intended hosts are necessary to cope with natural mutation accumulation and pervasive lateral gene flow. We present a generalizable strategy to preserve and constrain genetic information through the computational design of overlapping genes. Overlapping a sequence with an essential gene altered its fitness landscape and produced a constrained evolutionary path, even for synonymous mutations. Embedding a toxin gene in a gene of interest restricted its horizontal propagation. We further demonstrated a multiplex and scalable approach to build and test >7500 overlapping sequence designs, yielding functional yet highly divergent variants from natural homologs. This work enables deeper exploration of natural and engineered overlapping genes and facilitates enhanced genetic stability and biocontainment in emerging applications.

Protein-encoding information is stored in DNA as a series of trinucleotide codons. Because protein translation can occur in one of six coding frames, multiple proteins could in principle be produced from different frames of a single DNA sequence. Empirically, such overlapping genes are widely found in biology from bacteria to humans (*1*–*4*). Across microbial genomes, overlapping genes are estimated to make up almost one-third of all coding sequences (*5*). Although partial overlaps are more typical, many completely overlapped genes have been described (*1*, *6*) including an example where three different proteins are translated from separate frames of the same mRNA (*7*).

An important consequence of overlapping genes is that mutations will affect all protein products simultaneously. Mutations that would otherwise be neutral in one frame may no longer be permitted if they create deleterious mutations in another frame, which constitutes a mechanism to preserve sequence fidelity (*8*). In fact, biological systems experiencing very high mutation rates, such as viruses, tend to more frequently contain overlapping genes (*9*). Past synthetic efforts to maintain DNA sequence fidelity have focused on reducing background mutation rates (*10*, *11*), eliminating mutation-prone sequences (*12*), or increasing mutation surveillance and correction (*13*). To prevent escape of recombinant DNA into the wild, various biocontainment strategies have also been developed (*14*).

Inspired by naturally overlapping genes that safeguard against mutations, we devised a platform, Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS), to computationally design and experimentally test de novo overlapping genes (Fig. 1A). The overall computational objective is to identify co-encoding variants of two proteins of interest that share the same DNA sequence while minimizing disruptive residue changes in each protein sequence. Protein function can be disrupted by individual residue substitutions, insertions, or deletions, as well as by changes in long-range interactions between residue pairs. The CAMEOS algorithm addresses both considerations in two steps (Fig. 1B and fig. S1) (*15*). Briefly, a dynamic programming algorithm first generates a double-encoding solution that is optimal according to a hidden Markov model (HMM). High-performing suboptimal solutions are subsequently generated by a stochastic backtrace procedure. These HMM-derived solutions are used as seeds to the second step, in which pairwise long-range residue interactions

[1]Department of Systems Biology, Columbia University, New York, NY, USA. [2]Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA. [3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: hw2429@columbia.edu

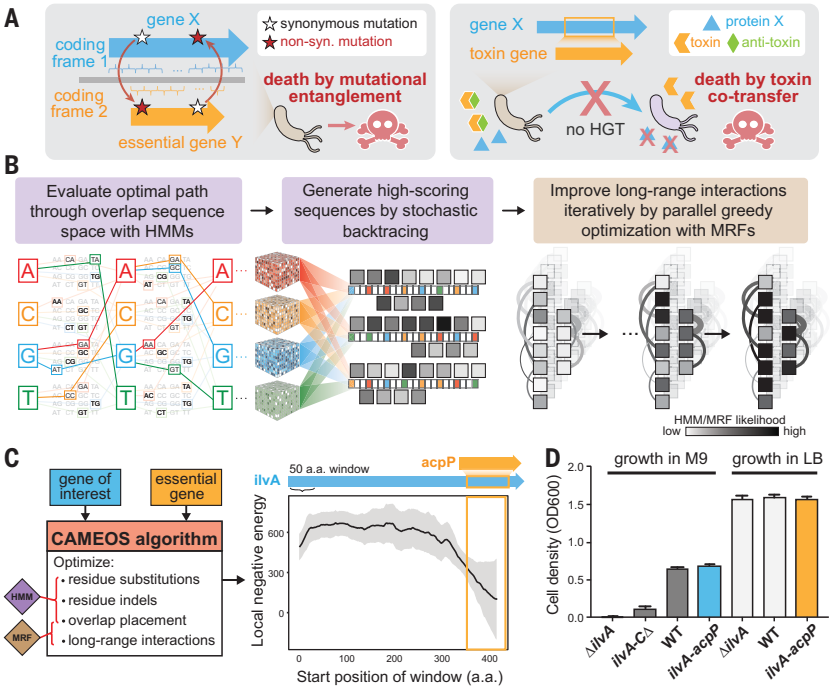

Fig. 1. CAMEOS platform for designing overlapping genes. (**A**) Schematic of mutational restriction or horizontal transfer confinement due to sequence entanglement of two genes. (**B**) The CAMEOS algorithm constructs a high-dimensional tensor (colored cubes) parameterizing the cost of paths (colored lines) through sequence space. These paths are then sampled probabilistically through a stochastic backtrace to form a population of sequences whose long-range interactions (gray arcs) are then optimized greedily and iteratively. (**C**) Parameters optimized by CAMEOS, with a schematic of co-encoded *ilvA* and *acpP* genes and the local negative energy of IlvA shown at the right. (**D**) Growth of a genomically encoded *ilvA-acpP* variant (IA-1) compared to that of control strains and wild-type (WT) cells after 14 hours in M9 minimal and LB rich media. The *ilvA*-CΔ strain is a Δ*ilvA* derivative with a plasmid expressing a C-terminal *ilvA* truncation variant at the overlap (residue 347). Data are means ± SEM from three independent biological replicates.

modeled by a Markov random field (MRF) (*16*, *17*) are iteratively optimized (fig. S2). We find that minimizing disruptions to long-range interactions is key for generating functional proteins (fig. S3 and table S1). Finally, synonymous mutations are made upstream of the overlap DNA region to optimize the Shine-Dalgarno sequence (*18*) for improved translation of the embedded gene.

We evaluated CAMEOS by designing and testing synthetic overlaps of essential and biosynthetic genes (Fig. 1C). Because certain protein regions such as intrinsically disordered regions may be more favorable for co-encoding (*19*), we explored different overlap positions and weights (fig. S4), including cases where one protein sequence was kept wild-type. Among 20 redesigned biosynthetic proteins, eight could rescue the growth of corresponding auxotrophic *Escherichia coli* bacteria in minimal media (fig. S5, A and B, and table S2). In one design, DE14, CAMEOS identified multiple optimal regions for overlapping, as shown by the ability to encode sequences similar to two essential proteins into separate regions of the functional cysteine biosynthesis gene *cysJ* (fig. S5C). The wild-type copy was knocked out in the cell to assess the co-encoded essential gene. In another design, DE2, we successfully generated a chromosomal deletion of the essential *acpP* gene in a strain containing overlapping *ilvA* and *acpP* genes. This DE2 construct encoded a redesigned protein sequence of IlvA (threonine deaminase, used in isoleucine biosynthesis) and a wild-type sequence of ACP (acyl-carrier protein, involved in essential fatty acid biosynthesis). We removed any plasmid-associated effects by genomically integrating this *ilvA-acpP* construct into a strain with deleted wild-type *ilvA* and *acpP* (fig. S6, A to C). The resulting strain (IA-1) exhibited isoleucine prototrophy at a wild-type level (Fig. 1D and fig. S6D), indicating functional activity of both biosynthetic and essential proteins.

To verify that sequence entanglement could restrict accumulation of mutations, we performed saturation mutagenesis on the genomic *ilvA-acpP* locus (Fig. 2A). Fitness effects for the first 30 overlapping codons of *ilvA* were assessed using oligo-recombineering and sequencing (*20*). Many mutations, especially in the beginning of the overlapping region, were found to incur a fitness defect (Fig. 2B and fig. S7). Although *ilvA* is not essential, 12.5% of *ilvA* mutations caused severe growth defects (decrease in growth rate by a factor of >10) (Fig. 2C). In contrast, mutagenesis of *ilvA-acpP* in a control strain (IA-2) that has an additional wild-type copy of *acpP* produced mutants with practically no growth defects (Fig. 2, B and C), which suggests that entanglement with the essential *acpP* gene renders the recoded *ilvA* gene sensitive to mutations. In the entangled sequence, a reduction in the degeneracy of codons was also observed, with 32% of synonymous codons in *ilvA* exhibiting high variability in their fitness impact because many were now deleterious (Fig. 2D). For example, the six leucine (L) codons had highly variable fitness effects across most of the overlap region analyzed (87%) (Fig. 2, B and D). Serial passaging of IA-1 and IA-2 showed that the overlap sequence remained unchanged in IA-1 after 150 generations, whereas mutations appeared in IA-2 by generation 50 (fig. S8). Together, these results demonstrate that a gene that overlaps with an essential gene is more evolutionarily constrained.

Because functional testing of in silico designs is a bottleneck, we devised a high-throughput synthesis and selection strategy to experimentally evaluate thousands of CAMEOS solutions. We used *cysJ*, a flavin sulfite reductase subunit (CysJ) for cysteine biosynthesis, and *infA*, the essential translation initiation factor–1 (IF1), as a test case by designing 7500 unique *cysJ-infA* overlapping solutions (Fig. 3A). CAMEOS designs were synthesized as a pool of 230–base pair

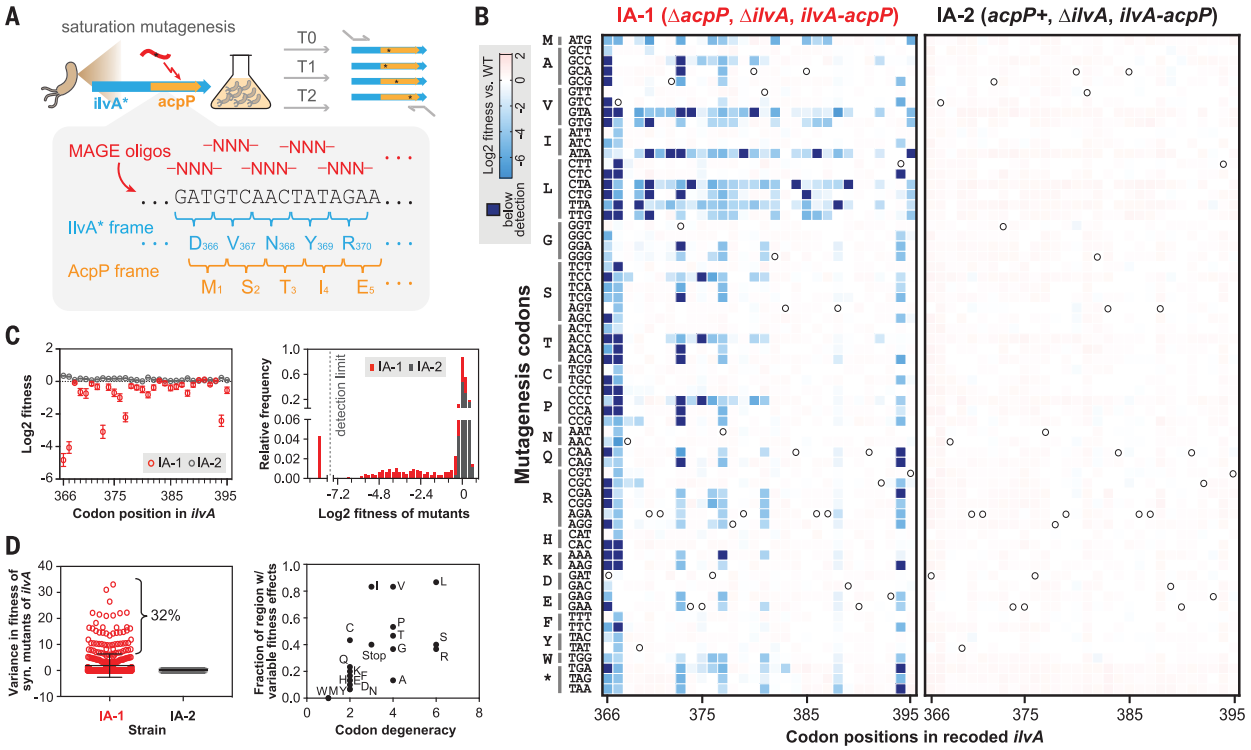

Fig. 2. Sequence entanglement alters the protein fitness landscape. (**A**) Saturation mutagenesis of the *ilvA-acpP* overlap region. (**B**) Fitness of all single-codon mutants of *ilvA-acpP* in strains IA-1 and IA-2. The *x* axis represents codon positions in the *ilvA*-coding frame; the *y* axis represents 64 mutagenized codons. White circles indicate wild-type *ilvA* codons. Heat map shows mutational fitness impact; dark blue indicates severe effects. (**C**) Left: Average fitness of all single codon mutants at each codon position in IA-1 (red) and IA-2 (gray) strains. Error bars denote SEM of the 64 codon mutants. Right: Distribution of fitness of all mutants in IA-1 and IA-2. (**D**) Left: Variance in fitness between synonymous mutants at each *ilvA* codon position. As shown, 32% of the codon substitutions in IA-1 have variances in fitness beyond the 95% confidence interval of IA-2 variances. Right: Fraction of the overlap region with highly variable fitness between synonymous codons for each amino acid plotted against its codon degeneracy. Amino acid abbreviations: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr; *, Stop.

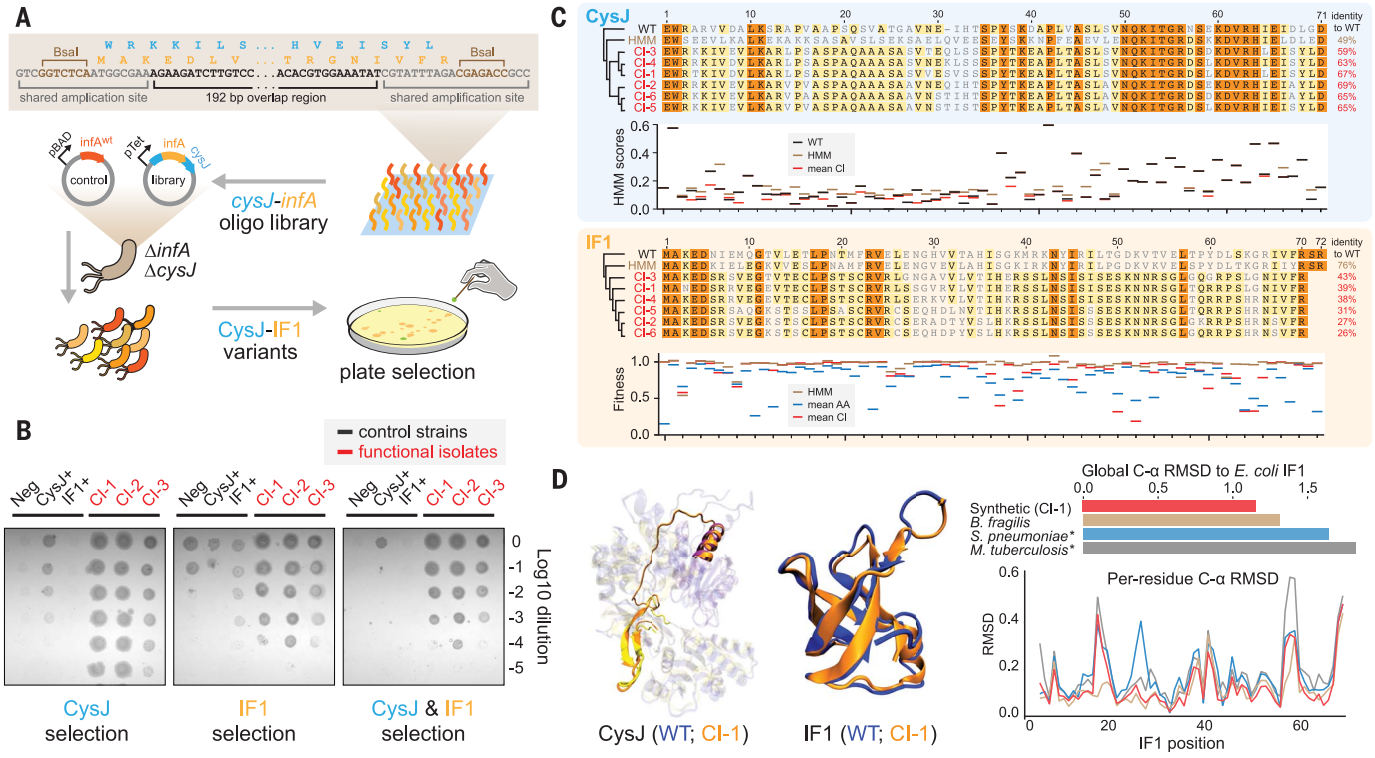

Fig. 3. High-throughput experimental evaluation of CAMEOS designs. (**A**) Diagram of selection platform to identify functional *cysJ-infA* variants. (**B**) Growth of *cysJ-infA* variants and controls under different plate selections. CI-1, CI-2, and CI-3 are different isolates encoding functional *cysJ-infA* variants. Control strains: Neg, cells containing pH9-cysJ^{neg}-infA^{neg} with nonsense mutations in both *cysJ* and *infA* reading frames; CysJ^{+}, cells expressing wild-type CysJ (from pH9-cysJ^{wt}); IF1^{+}, cells expressing wild-type IF1 (from pH9-cysJ-infA^{wt}). Cultures of each strain were spotted on selection plates over serial dilutions. (**C**) Multiple sequence alignment of IF1 and CysJ encoded by six functional *cysJ-infA* variants (CI-1 to CI-6). Colored shading represents different degrees of sequence identity: orange, 100%; yellow, >60%. Shown in panels below the sequence alignments are average fitnesses of IF1 mutants at every position based on saturation mutagenesis data (mean confidence interval) and single-residue substitution scores of CysJ based on the HMM model (mean confidence interval). WT, *E. coli* wild-type sequence; HMM, consensus sequence from HMM models; mean AA, average fitness of 20 amino acids. (**D**) Left: Structural modeling of CI-1 proteins shows concordance between predicted (orange) and crystal (CysJ, yellow/purple; IF1, blue) structures. The wild-type CysJ structure was generated by concatenating separately crystalized domains (yellow, purple). Right: Global (top) and per-residue (bottom) RMSD comparing IF1 from CI-1 with an ortholog model (*B. fragilis*) and other crystal structures (denoted by an asterisk: *S. pneumoniae*, *M. tuberculosis*) shows structural similarity to the *E. coli* IF1.

oligonucleotides (fig. S9 and table S6), cloned into a pH9-cysJ-infA^{entry} plasmid (fig. S10), and transformed into a Δ*cysJ-infA* strain (CI-Δ) for selection of functional CysJ and IF1 variants in MOPS dropout media (fig. S10) (*15*). Accordingly, CysJ and IF1 positive controls were only viable under their respective selective conditions (fig. S11). Plating of the variant library under double selection for both CysJ and IF1 function produced hundreds of colonies. We picked a number of colonies and reverified their phenotypes clonally (Fig. 3B) (*15*). Six unique sequences (CI-1 to CI-6) encoding different functional CysJ and IF1 variants were identified (Fig. 3C). Notably, all clones exhibited higher homology to CysJ (mean residue identity ~65%) than to IF1 (~34%) as well as lower homology to wild-type *E. coli* CysJ and IF1 than most natural variants (fig. S12). Surprisingly, functional IF1 variants contained residues that were deleterious as single-residue substitutions (*20*) (Fig. 3C). Analysis of the HMM likelihood scores of CysJ variants also revealed many single residues with predicted low fitness, which suggests that epistasis was exploited in our redesigned sequences. Structural modeling (*21*) of the CI-1 pair showed good structural alignment with wild-type CysJ and IF1, comparable to other natural orthologs (Fig. 3D). A large-scale computational analysis to overlap 119 essential with 49 biosynthetic *E. coli* proteins (*15*), which yielded ~5.8 million designs, showed that 531 of 5831 possible pairs (~9%) had pseudo-likelihood scores better than those of the experimentally verified *cysJ-infA* pairs (fig. S13). Accordingly, we estimate that 80% of these biosynthetic proteins could be encoded with at least one essential protein (fig. S14). Taken together, these results highlight that functional overlaps may exist for many gene pairs, which can be designed and evaluated at high throughput.

Finally, we hypothesized that sequence entanglement can also be used to generate biocontainment barriers that suppress unintended horizontal gene transfer (HGT). If a toxin gene is embedded in a gene of interest (GOI), recipients that lack the antitoxin would be killed when the co-acquired toxin is expressed (Fig. 1A). We thus tested designs of various bacterial toxins embedded in the *ilvA* gene and found *ilvA-ccdB* to be the best overlap pairing (Fig. 4, A and B, and fig. S15) (*15*). We then transformed conjugation-competent donors (D1, D2) expressing the antitoxin *ccdA* with plasmids carrying either *ilvA-ccdB* (T1) or *ilvA-ccdB^{stop}* (T2, containing a nonsense *ccdB* mutation) and incubated them with CcdA^{-} (R1) or CcdA^{+} (R2) recipients. As expected, CcdA^{+} recipients acquired *ilvA-ccdB* or *ilvA-ccdB^{stop}* plasmids from donors at similarly high efficiencies. In contrast, CcdA^{-} recipients acquired *ilvA-ccdB* at a much reduced frequency (by a factor of >2700) relative to the toxin-null *ilvA-ccdB^{stop}* control (Fig. 4C), thus demonstrating HGT suppression by CcdB-mediated killing. Interestingly, recipients that did acquire *ilvA-ccdB* remained auxotrophic for isoleucine (Fig. 4D). Accordingly, we found *ilvA-ccdB* mutations in these escapees that inactivated both *ilvA* and *ccdB* function. These results show that synthetic entanglement with a

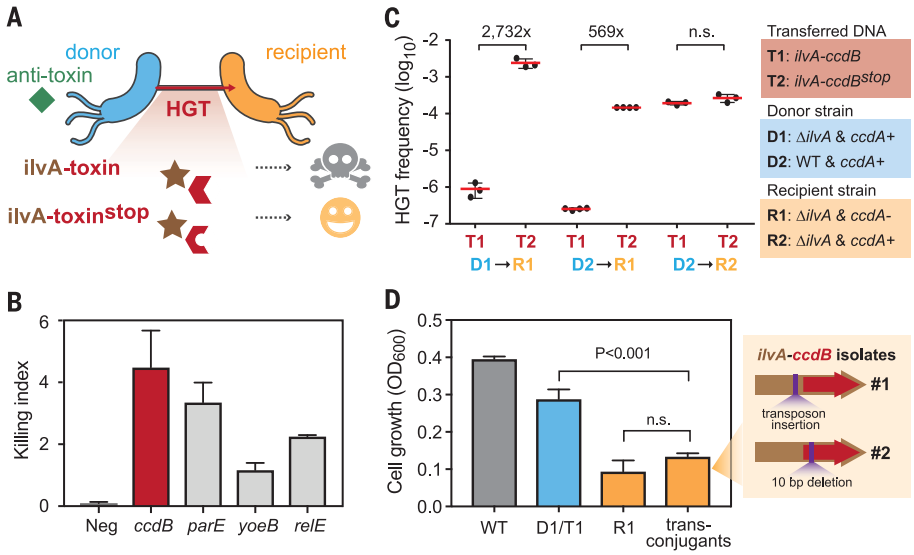

Fig. 4. Entanglement with a toxin limits HGT. (**A**) A toxin or an inactivated variant (toxin^{stop}) is overlapped with *ilvA* to assess HGT. (**B**) The killing index is shown for four *ilvA-toxin* gene pairs. Neg: parental strain without the *ilvA-toxin* construct. Data are means ± SEM from three independent experiments. (**C**) Efficiency of HGT of pHT plasmids (T1/T2) between different donor (D1/D2) and recipient (R1/R2) strains. Data are means ± SEM from three or four independent experiments. (**D**) Growth of R1 transconjugant isolates that acquired *ilvA-ccdB* (T1) after HGT, compared to the D1 donor strain (*ilvA*$^+$) and the R1 recipient strain (Δ*ilvA*) before HGT in M9 minimal media. Data are means ± SEM of growth measurements after 16 hours from three to five independent colonies and 28 R1 transconjugant isolates. Mutations identified in two R1 transconjugant isolates are illustrated at the right. Tukey's multiple comparison test was used to assess statistical significance (n.s., not significant).

toxin suppresses HGT and yields escapees that often carry nonfunctional GOI mutants.

This work describes the successful design of two full-length proteins into the same DNA under the constraints of overlap encoding. CAMEOS enables large leaps in protein space that are challenging to achieve through stepwise evolution. Our study goes beyond prior computational co-encoding efforts that used simple first-order BLOSUM substitution scores (*22*). Further improvements in MRF optimization (*23*) or explicit integration of protein structure information (*24*), along with higher-throughput gene synthesis methods (*25*), will facilitate the generation of longer overlapping genes with even higher performance. Better translation tuning may improve overlap designs, because suboptimal translation may be a failure mode. Ultimately, these advances can yield next-generation synthetic elements and circuits that will operate only in predefined settings, with greater robustness to mutations and over longer time scales.

REFERENCES AND NOTES

1. B. G. Barrell, G. M. Air, C. A. Hutchison 3rd, *Nature* **264**, 34–41 (1976).
2. C. A. Spencer, R. D. Gietz, R. B. Hodgetts, *Nature* **322**, 279–281 (1986).
3. I. B. Rogozin *et al.*, *Trends Genet.* **18**, 228–232 (2002).
4. C. R. Sanna, W.-H. Li, L. Zhang, *BMC Genomics* **9**, 169 (2008).
5. Z. I. Johnson, S. W. Chisholm, *Genome Res.* **14**, 2268–2272 (2004).
6. M. Mizokami *et al.*, *J. Mol. Evol.* **44** (suppl. 1), S83–S90 (1997).
7. J. Choi, Z. Xu, J. H. Ou, *Mol. Cell. Biol.* **23**, 1489–1497 (2003).
8. T. Miyata, T. Yasunaga, *Nature* **272**, 532–535 (1978).
9. E. Simon-Loriere, E. C. Holmes, I. Pagán, *Mol. Biol. Evol.* **30**, 1916–1928 (2013).
10. B. Csörgo, T. Fehér, E. Tímár, F. R. Blattner, G. Pósfai, *Microb. Cell Fact.* **11**, 11 (2012).
11. F. K. Balagaddé, L. You, C. L. Hansen, F. H. Arnold, S. R. Quake, *Science* **309**, 137–140 (2005).
12. B. R. Jack *et al.*, *ACS Synth. Biol.* **4**, 939–943 (2015).
13. A. Chavez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3669–3673 (2018).
14. J. W. Lee, C. T. Y. Chan, S. Slomovic, J. J. Collins, *Nat. Chem. Biol.* **14**, 530–537 (2018).
15. See supplementary materials.
16. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, *Proteins* **79**, 1061–1078 (2011).
17. H. Kamisetty, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
18. H. M. Salis, The ribosome binding site calculator. *Methods Enzymol.* **498**, 19–42 (2011).
19. C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, D. Karlin, *J. Virol.* **83**, 10719–10736 (2009).
20. E. D. Kelsic *et al.*, *Cell Syst.* **3**, 563–571.e6 (2016).
21. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **10**, 845–858 (2015).
22. V. Opuu, M. Silvert, T. Simonson, *Sci. Rep.* **7**, 15873 (2017).
23. V. Kolmogorov, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1568–1583 (2006).
24. B. Kuhlman *et al.*, *Science* **302**, 1364–1368 (2003).
25. C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, *Science* **359**, 343–347 (2018).

ACKNOWLEDGMENTS

We thank members of the Wang lab for advice and comments and D. Baker and S. Ovchinnikov for sharing the GREMLIN source code and data (available at https://gremlin2.bakerlab.org/preds.php?db=ECOLI). **Funding:** Supported by DARPA (W911NF-15-2-0065) and the Sloan Foundation (FR-2015-65795) (H.H.W.). **Author contributions:** T.B., H.-I.H., and H.H.W. developed the initial concept; T.B. and H.-I.H. performed experiments and analyzed the results under the supervision of H.H.W. All authors wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Sequencing data associated with this study are available in the National Center for Biotechnology Information Sequence Read Archive under PRJNA554669. The CAMEOS code is available at www.github.com/wanglabcumc/CAMEOS (commit a8d1ad3).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/365/6453/595/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S9
References (*26*–*31*)

28 September 2018; resubmitted 21 June 2019
Accepted 15 July 2019
10.1126/science.aav5477

Science

Synthetic sequence entanglement augments stability and containment of genetic information in cells

Tomasz Blazejewski, Hsing-I Ho and Harris H. Wang

Science **365** (6453), 595-598.
DOI: 10.1126/science.aav5477

Overlapping genes for synthetic biology

Overlapping genes yield multiple distinct proteins when translated in alternative reading frames of the same nucleotide sequence. Blazejewski *et al.* developed a computational algorithm to predict de novo sequence entanglement and experimentally generated functional synthetic overlapping genes. When a sequence of interest was co-encoded with an essential gene in a living bacterium, its evolutionary stability substantially increased. When a gene of interest was synthetically overlapped with a toxin gene, its horizontal gene transfer frequency between bacteria was strongly suppressed. This generalizable strategy for designing, building, and testing overlapping genes helps stabilize vertical gene evolution and restrict horizontal gene flow.

Science, this issue p. 595

science.sciencemag.org/content/365/6453/595/suppl/DC1

Supplementary Materials for

Synthetic sequence entanglement augments stability and containment of genetic information in cells

Tomasz Blazejewski, Hsing-I Ho, Harris H. Wang*

*Corresponding author. Email: hw2429@columbia.edu

Published 9 August 2019, *Science* **365**, 595 (2019)
DOI: 10.1126/science.aav5477

This PDF file includes:
Materials and Methods
Supplementary Text
Figs. S1 to S15
References

Other supplementary material for this manuscript includes:
Tables S1 to S9 (combined Excel file)

Materials and Methods

Strains and culturing conditions

All *E. coli* strains used in the study are listed in **Table S3** and are derived from the BW25113 parental background. Cells are grown in either LB rich media or M9 (prepared from BD Difco M9 Minimal Salts 5x; 1x M9 medium was supplemented with 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and 0.4% glucose) at 37 °C unless stated otherwise. For antibiotic selection, induction, or growth experiments, the concentrations used are: Carbenicillin (Carb) 50 μg/ml, Chloramphenicol (Cam) 20 μg/ml, Kanamycin (Kan) 50 μg/ml, Spectinomycin (Spec) 50 μg/ml, Bleocin (5 μg/ml), Anhydrotetracycline (aTc) 100 ng/ml.

Plasmids used and generated

Double encoded sequences were synthesized either by IDT as gBlocks or by Gen9 as GeneBytes and cloned under pTetO control in the pH9 plasmid. Plasmid pH9 was constructed by joining spectinomycin resistance cassette, p15A ori, *tetR* and pTetO promoter region from pdCas9-bacteria (Addgene #44249) through isothermal assembly. To construct pH9-cysJ^{wt} plasmid, the wild-type *cysJ* sequence was amplified from the *E. coli* K-12 genome and cloned into pH9 plasmid behind the pTetO promoter. pH9-cysJ-infA^{wt} was generated by swapping the designated *cysJ* region (residues 196-253) with overlapping wild-type *infA* sequence. pH9-cysJ^{neg}-infA^{neg} was constructed by introducing multiple stop codons in both *infA* and *cysJ* reading frames from pH9-cysJ-infA^{wt}. The *cysJ-infA* overlapping sequence was also replaced by short sequences flanked by BsaI cutting sites and optimized RBS200 sequence to create a *cysJ* entry vector (pH9-cysJ-infA^{entry}) for subsequent *cysJ-infA* library construction. pBAD-infA was constructed by assembling wild-type *infA* under pBAD and AraC control, Sc101 ori, and Bleocin resistance cassette. RK2-mobilizable plasmids pHT-T1 and pHT-T2, encoding *ilvA-ccdB*^{wt} and *ilvA-ccdB*^{stop} were generated by modifying the pH9-ilvA-ccdB^{wt} and pH9-ilvA-ccdB^{stop} plasmids by replacing *tetR* with an origin of transfer (oriT) sequence from the RP4/RK2 conjugative plasmid to enable mobilization by RK2-mediated conjugation, which also leads to constitutive expression of the *ilvA-ccdB* double-encoded gene. The pH-ccdA plasmid that expresses the anti-toxin *ccdA* was constructed to include a Cam resistance gene, *tetR*, and a constitutively expressing *ccdA* (under the BBa J23119 promoter), and a ColE1 ori. Most plasmids are generated (if not otherwise noted) by isothermal assembly. All plasmids are listed in **Table S4**.

Functional assessment of double-encoded genes

We first explored the ability of CAMEOS to yield functional proteins even in the absence of any overlap encoding to validate general model performance. Protein variants with a range of pseudolikelihood scores were assessed. We used amino acid biosynthesis and antibiotic resistance genes as test cases and synthesized variants without gene overlap. Biosynthetic gene variants (*trpE*, *cysJ*, *ilvA*) were tested in corresponding auxotrophic *E. coli* strains for the ability to rescue growth in minimal media. Chloramphenicol resistance gene variants were assessed for cell viability under antibiotic selection. For double-encoded candidates, constructs expressing different biosynthetic genes were transformed to corresponding auxotrophic strains (Δ*ilvA*, Δ*cysJ*, and Δ*trpE*), as shown in **fig. S5A**. Experiments to obtain growth curves were conducted as follows: Bacteria were grown in LB with antibiotic selection overnight at 37 °C, diluted 1:100 with fresh medium, grown for 1 hour, and then induced with aTc for 2 hours for expression of the double-encoded biosynthetic gene. After induction, cells were collected by centrifugation and washed with PBS twice before

diluted 1:40 in M9 minimal medium containing Spec (50 μg/ml) and aTc (100 ng/ml). Cell growth at 37 °C was measured continuously using a platereader (BioTek Synergy H1) at optical density of 600nm for 15-20 hours in 20-minute intervals.

Genomic knockout of essential genes

Genomic deletion of essential genes (shown in **fig. S5A**) was carried out as following. Cells with plasmids containing recoded essential genes were made electrocompetent and transformed with pKD46 for recombineering. Subsequently, they were grown and λ-Red proteins were induced prior to transformation with a double-stranded linear cassette targeting the wild-type essential gene for knockout. Knockout cassettes were generated using a Cam resistance cassette as the template and primers with 50-bp overhang sequences homologous to the targeted genomic locus. Primers used were designed for complete deletion of the coding sequence. After electroporation of the knockout cassette, cells were recovered in SOC (NEB) with aTc (to induce the recoded essential gene) for 1 hour before plating on LB with Cam, Spec, and aTc at 30 °C. Colony PCR and Sanger sequencing was performed on resulting colony isolates to confirm the deletion.

Construction of IA-1 and IA-2 strains

To remove any effects from plasmid copy number variation of the *ilvA-acpP* construct, the entangled *ilvA-acpP* sequence was directly integrated into the *E.coli* genome through the clonetegration system *(26)* (shown in **fig. S6A**). DE2 and the *tetR* regulon region was cloned between multiple cloning sites in pOSIP-CT and transformed into Δ*ilvA*. Chromosomal integration of the construct was induced by heat shock and successful integrants are selected on LB-Cam plates. Correct genomic insertion was confirmed by colony PCR and Sanger sequencing. Extra DNA sequence from the integrative plasmid backbone was excised from the chromosome through flipase induction as described *(26)*. Cells with DE2 insertions were further transformed with pKD46 to knockout the wild-type *acpP* as described above. We were unable to obtain *acpP* knockouts after a few attempts in contrast to successful deletion of *acpP* in a strain containing a plasmid copy of DE2. We hypothesized that insufficient expression of DE2 from a single genomic copy (thus leading to an ACP limitation in the cell) may be the cause of this failure to delete the endogenous *acpP* gene, especially since ACP is the most abundant protein in *E. coli*. Therefore, we attempted to modify the internal RBS upstream of the embedded *acpP* to increase the translation strength using MAGE. By sequentially performing MAGE and *acpP* knockout, we successfully obtained clones with knockout of the wild-type *acpP* gene (**fig. S6B-C**), suggesting that indeed translation tuning may be important during some scenarios of double-encoding optimizations. The pKD46 plasmid was subsequently cured to yield the strain, IA-1. The expected growth phenotype of IA-1 was verified in minimal medium (**Fig. 1D, fig. S6D**). IA-2 was generated by re-introducing the wild-type *acpP* gene into IA-1 by recombineering. We screened IA-2 isolates for Cam sensitivity and confirmed genomic restoration of wild-type *acpP* by colony PCR and Sanger sequencing.

MAGE mutagenesis on IA-1 and IA-2 strains

We performed saturation mutagenesis on the IA-1 and IA-2 strains using a version of MAGE *(27)* with higher efficiency. Tiling 90-bp single-stranded mutagenesis oligos (**Table S5**) were designed to target a 30-residue window of the *ilvA-acpP* genomic locus by flanking homology around consecutive degenerate trinucleotides (NNN) in the middle of each oligo. Oligos were designed to target the first 30 codons of the recoded *ilvA* reading frame of *ilvA-acpP*. In all,

30 pools of degenerate oligos were synthesized (Integrated DNA Technologies). In each oligo set, the last 5 nucleotides at the 3' and 5' ends were phosphorothioated for improved MAGE efficiency. Three successive rounds of MAGE were performed on each strain to elevate the rate of resulting mutagenesis. In each round, cells were electroporated with a pool of the 30 oligos (15 μM total) after λ-Red induction. The resulting mutant population was grown and sampled at various time points for deep sequencing. In detail, cells after the last recovery step were washed with PBS twice and subjected to growth in M9 supplemented with 0.5 μM isoleucine and aTc at 30 °C. At subsequent time points, 1 ml of cell culture was collected, and the remaining culture was diluted 1:100 with fresh medium to re-grow.

Fitness measurement of ilvA-acpP mutants

Cell culture was collected at 0 hours, 18 hours, 42 hours, and 64 hours after MAGE mutagenesis. Genomic DNA was extracted by prepGEM kit (Zygem) and the *ilvA-acpP* overlapping region was amplified by PCR. A second PCR was performed to add adaptor sequences and barcodes compatible with Illumina sequencing kit (Illumina, Nextseq MO300). Reads were filtered for quality and variants were identified. Variants (>Q30) with mutations in more than one region of the *ilvA-acpP* sequence were excluded. Fitness of each variant was calculated by taking the relative abundance ratio between time points and normalizing against the relative abundance of the wildtype sequence. The median value of the calculated ratios across time points were used. Fitness data for IA-1 and IA-2 are provided in **Table S7-S8**.

Design and construction of the cysJ-infA library

7,500 cysJ-infA recoded variants were synthesized on a DNA microarray (Agilent Technology). In the general case, sequences were generated using HMM optimization followed by greedy optimization of MRF pseudolikelihoods. The relative importance of the *infA* and *cysJ* pseudolikelihoods for the optimization objective was controlled by a CysJ weight factor between 0.0-1.0 that was randomly generated from a uniform distribution for each non-control sequence. 4,063 of the 7,500 sequences were optimized for 1000 steps, and 1,924 sequences were chosen from earlier points (600 steps, 400 steps, 200 steps, and 50 steps) to evaluate the role of pseudolikelihood optimization. Similarly, 496 sequences were initialized with wild-type IF1 amino acid sequences and synthesized over the course of *in silico* optimization. Several other control sequences were also synthesized: 519 sequences that were optimized only through HMM optimization; 250 sequences with wild-type IF1 amino acid sequence; 250 sequences with wild-type CysJ amino acid. RBS designs were derived by evaluating all synonymous variants of *cysJ* 27 bp upstream of the *infA* start site followed by manual refinement with the RBS Calculator.

A previous saturation mutagenesis study of IF1 (*19*) revealed that truncation of the final two amino acids of the IF1 protein did not impact cell fitness. In our designs, these two residues were removed to reduce the number of nucleotides required for synthesis. Similarly, the stop codon for the *infA* gene was placed downstream of the insertion point on the pH9-cysJ-infA^{entry} plasmid. The first 8 bases of the overlap region (ATGGCGAA), as well as the last 10 bases (CGTATTTAGA), were held constant during optimization, such that these flanking sequences could be used for PCR amplification during library construction. Before making these sequence-level changes, we verified experimentally that the amino acid substitutions had no effect on the activity of *E. coli* CysJ protein. The 230bp oligomers were amplified by high fidelity polymerase (Q5 DNA polymerase, NEB) with minimum cycles to avoid over-amplification. PCR products were purified, digested with BsaI, and ligated with BsaI-treated pH9-cysJ-infA^{entry} vector. Ligation

solution was transformed to commercial competent cells (MegaX DH10B, Invitrogen) following the manufacturer's protocol. Cells after recovery were plated at the appropriate density on LB-Spec plates to estimate cloning efficiency and library complexity. A library with at least 40x average coverage of each member was generated. The remaining cells were expanded in 50 ml of LB with selection overnight. The *cysJ-infA* library was extracted from the overnight culture with Midi-prep DNA extraction kit (Zymo Research) for subsequent testing.

Selection of functional cysJ-infA variants

Verification of functional *infA* designs requires an inducible expression platform of the essential IF1 protein. We transformed pBAD-infA and pKD46 to a Δ*cysJ* strain. Wild-type *infA* was deleted by recombineering under episomal IF1 induction to yield the CI-Δ strain. CI-Δ was confirmed to be not viable unless IF1 was induced (**fig. S10A**). CI-Δ was then transformed with the 7,500 member *cysJ-infA* library. Transformants were grown overnight in LB with arabinose induction and antibiotic selection. Overnight culture was diluted 1:100 with MOPS media (TEKNOVA) with no arabinose induction and grown for 1 hour at 37 °C. Expression of CysJ and IF1 variants were induced with aTc for another 3 hours before plating on selective conditions. Cells were plated at the density of $6x10^4$ cells per 100 mm plate. Selective plates are made with MOPS defined medium with 1% glycerol as carbon source, and Spec and Bleocin for plasmid maintenance. Adjustment was made for different selective conditions as follows. CysJ selection: MOPS without cysteine, aTc, arabinose; IF1 selection: MOPS, aTc; CysJ/IF1 dual-selection: MOPS without cysteine, aTc. After 3 days of incubation at 30 °C, grown colonies were harvested either as a pool by scraping the whole plate or picked individually. The library was extracted by Mini-prep kit (Qiagen). For the pooled library, we amplified the designed overlapping region and quantified the abundance of each variant by deep sequencing (Illumina, Nextseq HO300). For clones identified from CysJ/IF1 selection, isolates were individually transformed back to the parental CI-Δ strain and revalidated for growth phenotype under CysJ/IF1 selection. Out of 36 clones identified from the initial screen, seven (CI-1 to CI-6, CI-1 was identified twice) showed consistent growth under dual-selection. Two out of the six variants (CI-2 and CI-5) contained additional point mutations beyond their original designs, which were likely generated during gene synthesis or assembly, highlighting that additional variants were explored over the course of the experiment.

Structural modeling of the identified CysJ and IF1 variants

CysJ and IF1 sequences were modeled using the Phyre2 online tool *(20)*. The tool was run in "intensive" modeling mode. Proteins were analyzed and visualized using VMD. Model overlay parameters were obtained through "measure fit" commands over aligned residues.

Validation of ilvA-toxin constructs and quantification of degree of horizontal gene transfer

We calculated the "killing index" of every *ilvA*-toxin construct by quantifying the activity of *ilvA* and toxins (*relE*, *parE*, *yoeB*, and *ccdB*) separately. In each *ilvA-toxin* pair, additional nonsense mutations were introduced in the toxin reading frame ($toxin^{stop}$) without disrupting the corresponding residues in the *ilvA* frame (i.e. using synonymous mutations) (**Fig. 4A**). The function of each recoded *ilvA* variant could thus be tested without the confounding impact of the co-encoded toxin. Each construct was then transformed to the *ilvA* knockout strain and their growth was measured in minimal media (**fig. S15A**). To then examine the function of the overlapping toxin, we reverted the nonsense mutations in the toxin back to the wild-type sequence ($toxin^{wt}$) and

tested each construct in wild-type cells. Changes in cell growth were used to determine cellular toxicity due to the toxin (**fig. S15B**). The *ilvA* activity was determined based on how well *ilvA-toxin*^{stop} constructs can rescue the growth phenotype of a Δ*ilvA* strain in minimal media. The toxin activity was determined by measuring their growth inhibition in wild-type cells when overexpressed on a plasmid. Killing index was calculated from OD600 cell density data after 16-hour of growth in LB and M9 media of cells containing *ilvA-toxin*^{stop} or *ilvA-toxin*^{wt} and wild-type cells using the following equation:

$$\mathrm{OD600}_{\mathrm{M9}}\left[\frac{ilvA\text{-}toxin^{stop}}{\mathrm{WT}}\right] \Big/ \mathrm{OD600}_{\mathrm{LB}}\left[\frac{ilvA\text{-}toxin^{wt}}{\mathrm{WT}}\right]$$

For quantification of horizontal gene transfer, donor strains were all conjugated with RK2/RP4 plasmid to mediate subsequent DNA transfer. Donor and recipient strains were grown overnight in rich medium with antibiotics selection. Strains were collected, washed and re-suspended in 1/10th of the original volume of PBS, and mixed at 1:1 ratio. The cell mixture was spotted on LB plates at 37 °C for 2 hours, and subsequently collected and plated at multiple densities under different antibiotic selection. For mixture of strains derived from Δ*ilvA*, recipients were labeled with Bleocin resistance gene to distinguish from the donors. We used Bleo or Bleo+Spec dual selection to quantify conjugation events. For mixture of Δ*ilvA* and wildtype derived strains, all recipient strains have Kan resistance gene, but only those receiving the mobilized plasmid can grow under Kan and Spec selection. Hence, the ratio of colony numbers grown on dual versus single selection plates can be calculated to infer the efficiency of horizontal gene transfer.

Data sources

In experiments involving cysJ and ilvA, publicly available HMMs and multiple sequence alignments (MSA) were used. For HMMs, we made use of TIGRfam and Pfam models. MRFs were trained based on MSAs available from the Gremlin2 project https://gremlin2.bakerlab.org/preds.php?db=ECOLI. MRFs were trained on these MSAs using Gremlin code kindly shared by the Baker Lab (University of Washington, for request http://openseq.org/gremlin.php). Subsequent computational analyses used InterPro as a source of multiple sequence alignments (accessed on 4/24/2019) and CCMPred as a MRF optimization algorithm.

Code availability

CAMEOS is implemented in Julia and Python. The code needed to run the HMM and MRF optimization process is available at www.github.com/wanglabcumc/CAMEOS. Commit version: a8d1ad3.

Supplementary Text

Overview of the CAMEOS computational algorithm

To design overlapping sequences, the Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS) algorithm proceeds in two steps (**fig S1**). First, we seek to tractably and efficiently generate double-encoding solutions that exhibit high homology to target sequences: this is achieved by a first-order optimization process that is parameterized by Hidden Markov Models (HMM). In order to generate a diverse set of possible solutions, we then perform probabilistic backtraces to generate a stochastic population of relatively high-quality solutions. Second, we use Markov Random Field (MRF) models to optimize long-range interactions between residues, which are not captured by HMM models and yet are essential for protein functions, in the HMM-derived solutions to yield final refined overlapping solutions. Each of these steps is described in greater detail below.

First-order optimization: HMM parametrization and stochastic backtrace

The first step of CAMEOS operates on tetra-nucleotides units, each of which fully encapsulates two codons in overlapping frames (**fig. S1A-B**). The optimization goal is thus to find the optimal set of tetra-nucleotides suitable for both co-encoded proteins. This is a first-order optimization as we do not yet consider residue-residue epistatic interactions (incorporated in the second step of CAMEOS). Thus in this step, we assume that the optimal residue at each position can be determined independently of the rest of the protein sequence. Despite this simplifying assumption, optimization is still non-trivial as each tetra-nucleotide overlaps with other tetra-nucleotides through its first/fourth position so that globally optimal first-order decisions must balance the score of a current tetra-nucleotide with the potential scores of downstream tetra-nucleotides.

If we simplify downstream effects by assuming that the first/fourth bases for a given tetra-nucleotide are fixed, optimization of the entire tetra-nucleotide can be achieved by using a HMM to score residue translations produced by each of the sixteen di-nucleotide possibilities for the second/third base (**fig. S1B**). The log-likelihood values of every possible tetra-nucleotide can be easily computed from a Hidden Markov Model (HMM) of the corresponding protein family (**fig. S1C**). Since tetra-nucleotides are not independent, naïve optimization would require considering every possible combination of first/fourth bases across all tetra-nucleotides; this would lead to an exponential number of possible combinations of sequences of a given length (**fig. S1D**). However, we observe that optimization decisions in the second tetra-nucleotide unit depend only on the fourth base of the first tetra-nucleotide unit. An optimal 7-letter sequence must extend earlier optimal tetra-nucleotides whose sole distinguishing characteristics in terms of downstream optimization decisions is the score up to this point and its final nucleotide: “A”, “C”, “G”, or “T”. This is true for all subsequent positions: an optimal subsequence of length n is obtained by optimally extending an optimal subsequence of length n-3 distinguished by its earlier scores and its terminal letter, “A”, “C”, “G”, or “T” (**fig. S1E**). The optimization algorithm therefore finds optimal subsequences ending in “A”, “C”, “G”, or “T” for each position by extending earlier subsequences ending with these nucleotides. If the current tetra-nucleotides is at the end of the entire sequence, we can determine the globally optimal sequence by comparing the score of optimal subsequences across all terminal letters (**fig. S1F**).

Internally, optimal sequences are constructed as paths between scored subsequences distinguished by their terminal nucleotide. If we wish to generate a set of high-scoring but sub-optimal set of sequences for experimental validation (i.e. instead of always connecting maximally

scoring subsequences), we can slightly alter the algorithm to stochastically incorporate high-scoring but sub-optimal subsequences into the sequence generation procedure. This procedure, which we refer to as a "stochastic backtrace" (**fig. S1G**), allows a significantly larger sequence space to be explored, which is useful in cases where we wish to evaluate many sequence variants.

In practice, this step involves starting at the initial starting point and considering the scores of all feasible transitions from each state parameterization. We can either choose the next state (and corresponding sub-sequence to add to the solution) probabilistically at each step or deterministically most of the time with an occasional stochastic switch to a probabilistic mode. This mostly deterministic strategy can still generate a diversity of solutions as suboptimal decisions at given points (such as adding an unnecessary insertion) can lead to very different downstream optima. In general, the probabilistic backtrace is therefore done by making deterministic decisions with probability = 0.9, meaning that roughly 10% of the time, a stochastic decision is taken.

While other stochastic backtrace algorithms could be considered, the strategy implemented here is able to generate a diverse set of solutions. This diversity is obtained due to the large possible set of transitions at each position (i.e. transitions across nucleotides, states, and positions) and inherent flexibility in possible solutions. Indeed, many overlapping sequence solutions appear to have similar log-likelihoods based on HMM parameterizations. We provide an annotated iPython notebook (https://www.github.com/wanglabcumc/CAMEOS/guide.ipynb) to illustrate the basic premise of this part of the algorithm. This code is a Python implementation of the generalized CAMEOS code, which is written in Julia, but is simpler to read and understand at the cost of slower speed.

Second-order optimization: greedy optimization through MRFs

Long-range interactions between residues are vital to protein structure and function. Sequences optimized only through first-order considerations are likely to miss essential stabilizing contacts. The first-order optimization is therefore used to seed the second optimization step for long-range interactions as parameterized by a MRF. From a seed population of thousands of stochastic backtrace solutions derived from the first CAMEOS step, we use a greedy optimization procedure in the second CAMEOS step and iterate for a fixed number of optimizations (**fig. S1H**).

The optimization procedure is very closely related to iterated conditional modes (ICM), a general strategy for optimizing trained MRFs. The algorithm is greedy: at each optimization step, for each individual sequence in the population, we select three contiguous nucleotides (i.e. tri-nucleotides) from a random position along the length of the overlapping sequence (disregarding codon positioning), and consider all possible tri-nucleotides (not resulting in a stop codon) that could replace the current one. The score of the tri-nucleotide, which is assessed as a sum of MRF pseudolikelihoods across both frames, is determined assuming the sequence of the rest of the sequence remains fixed. The sum of pseudolikelihoods can also be weighted, in order to favor one protein's scores over the other. At each step, the maximal scoring tri-nucleotide is used to continue the optimization process through subsequent iterations.

The difference between this strategy and standard ICM is that the position in the sequence to be optimized is chosen stochastically. This is done because the residue identities and therefore "node values" of the MRF are linked between sequences, meaning that individual node values cannot be modified in isolation. In theory, optimization can be performed deterministically by considering every tri-nucleotide at every position of the sequence, but this would repeat calculations (e.g. as there is a defined set of amino acids that can result from a codon beginning in

“A”). We therefore choose positions at random and empirically observe convergence upon optimization.

Other key CAMEOS features

Insertions: The algorithm described above does not consider residue insertions, which may increase the flexibility of the sequence generation procedure and yield in higher-quality sequences. We thus incorporate insertions by scoring tetra-nucleotides across every possible alignment between the two proteins, an operation requiring O(mn) computation, where m and n are the lengths of each protein. The probability of an insertion is defined by the same protein-specific HMM, and optimal subsequences considered for extension are now defined not only by their terminal letter but also by their insertion state (true/false) and position in both proteins. We note that though HMMs can incorporate insertions and deletions, MRFs are only able to consider insertions/deletions already present in the multiple sequence alignment they were trained on. In cases where an HMM has generated an insertion, we consider the HMM solution to have a better model of this likelihood than its corresponding MRF, and the sequence at this inserted position is not allowed to change over the course of greedy optimization rounds.

Other frames: While we have only illustrated co-encoding the +1 frame, our algorithm is applicable to all frames. Consider two proteins D_1 and D_2. In a +2 encoding, the first nucleotide of D_2's codons immediately precedes the first nucleotide of the D_1 codon. This is equivalent to a +1 encoding where D_1 and D_2 are switched; this is how a +2 encoding is optimized in our model. This observation is also true of -1 and -2 encodings, which also share the fundamental property that codons can be optimized as tetra-nucleotides, and are therefore amenable to the same optimization algorithm. The -3 frame can be optimized trivially: codons perfectly overlap as reverse complements in both frames, so locally optimal codon decisions are globally optimal according to a first-order model and no recurrence is required.

Optimal regions of overlap: We found that an important factor to consider when designing overlapping sequences is the "local energy" of a protein across a given region. This metric simply sums node weights (corresponding to metrics of conservation for amino acids at a given position) and edge weights (constituting the long-range interactions at each position) originating from this region. If a region captures the entire length of a protein, then this "local energy" corresponds exactly to the MRF "energy" of the protein itself. We note that due to a missing intractable partition function, comparisons of energy values between proteins are not meaningful. However, a visualization of the energy of a single target protein across windows of length equal to that of the query protein being embedded can qualitatively reveal regions in the target protein that are expected to be amenable for incorporation of the query protein. Given that the density and importance of long-range interactions can vary across the length of a protein, it may often be advantageous to target specific regions of a protein for overlap encoding. In cases where specific regions are targeted for overlap, we simply penalize the cost of incorporating residues outside of this range with a very large constant. During optimization, residues outside of the desired range are not incorporated due to their high cost.

Generality of sequence entanglement using CAMEOS

To assess factors impacting the applicability of CAMEOS to further protein pairs, we performed a large-scale computational survey of the algorithm’s performance. Beginning with a list of 213 essential genes and 71 biosynthetic genes, we queried InterPro *(28)* to automatically identify protein families, selecting the most specific subfamily in cases of multiple hits. Following

guidelines for MRF training using GREMLIN, we used the ratio of non-redundant sequences in this alignment to the square root of the length of the input protein as a threshold for family inclusion. A conservative threshold of 250 was used to select 49 biosynthetic proteins. Given the large number of essential genes, we selected 119 essential genes with a more conservative threshold of 500 (**Table S9**). Sequences belonging to the targeted protein family were downloaded from InterPro and aligned using FAMSA *(29)*. Outliers in the multiple sequence alignment were then excluded using OD-seq *(30)* and a standard deviation threshold of 2. Sequences were then aligned again using FAMSA. Multiple sequence alignment positions in which less than 50% of entries were aligned amino acids were excluded. To speed up the MRF training step of our computational analysis, we used *CCMPred (31)*, a tool that applies the same algorithm as GREMLIN but with significantly improved speed due to more efficient computation. HMM models were trained on the same alignments as the corresponding MRFs using HMMER.

Using our CAMEOS code (at https://www.github.com/wanglabcumc/CAMEOS/), we tested co-encoding of all 5,831 pairwise combinations of essential and biosynthetic genes. This analysis, which involved optimization of ~5.8 million sequences, required approximately 3 days of computational analysis across 384 CPU threads. Local energy minima for various embedding lengths were estimated based on a random subset of 2,000 sequences for each target alignment. For each pair of MRFs, we initialized 500 solutions and optimized sequences for 1000 steps. As we were only interested in top-scoring variants, we culled sequences with worse-than-median scores at step 250 and step 500. In order to compare optimized sequences with natural variants, we compared final pseudolikelihood values with the distribution of pseudolikelihood scores of sequences in the multiple sequence alignment used to train our Markov Random Field models. These were found to follow approximately Gaussian distributions, but in order to further improve the robustness of our analysis, we used the median and median absolute deviation (MAD) with Gaussian scale factor as an estimator of the location (μ) and scale parameters (σ), of our observed distributions.

For each double-encoding solution, we determine the sum of the Gaussian distributions for each protein pair (essential and biosynthetic). According to standard definitions, the sum of Gaussian distributions for the two protein pairing, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ have a mean of $\mu_1 + \mu_2$ and a standard deviation of $\sqrt{\sigma_1^2 + \sigma_2^2}$. We assessed the z-score of our solutions according to this distribution with the expectation that sequences closer to the median are more likely to be functional. A few solutions generated with different z-scores were depicted in **fig. S14**. In general, codons with high conservation tend to be maintained or altered with similar amino acids after optimization, suggesting the effectiveness of our algorithm. We find that a dominant factor influencing the z-score of tested pairs was the length of encoded sequence overlap (R^2=0.56, **fig. S13B**). The axes of the heatmap in **fig S13A** are sorted by length, with the top-right portion of the heatmap reflecting high z-scores in the case of long proteins embedding in longer proteins.

We select a z-score of 3.0, similar to the minimal z-score across all synthesized array variants of 2.84 that was achieved in our original analysis, but more conservative than the z-score of 3.635 possessed by the functional variant CI-4. We find that this z-score is in the 9th percentile of all tested pairs, with 531 gene pairs exhibiting z-scores that outperform this mark. We note that the minimal z-score of 3.635 across functional *infA-cysJ* design pairs would be outperformed by 794 gene pairs (~13.6% of all tested pairs). In the case of a more stringent z-score of 2.0, we find that 235 gene pairs (~4.0%) would meet the threshold.

This computational simulation, benchmarked against a double-encoding solution that has been demonstrated to function, suggests that hundreds more gene pairs could be suitably targeted

by our algorithm. We further find that with the use of a z-score threshold of 3.0, 41 of 49 biosynthetic genes could be encoded with at least one essential gene, and that 44 of 119 tested essential genes could be encoded with at least one biosynthetic gene.

Interested in features that might explain the residual spread after accounting for length, we directly examined variations in the residuals by protein family and found that deviations appeared to be the result of protein family-specific variations (**fig. S13D, E**). We observe first that the value of the residual for any gene pair (*S*, *B*) is tightly related to the sum of means across all gene pairs containing *S* and all gene pairs containing *B* (**fig. S13F**). We further observed that residual values of +1 and +2 encoding for identical gene pairs were also tightly correlated (**fig. S13G**). These observations suggest that double-encoding success is largely dependent on fundamental features of the protein families being encoded as opposed to complex nucleotide and codon-specific interactions between gene pairs.

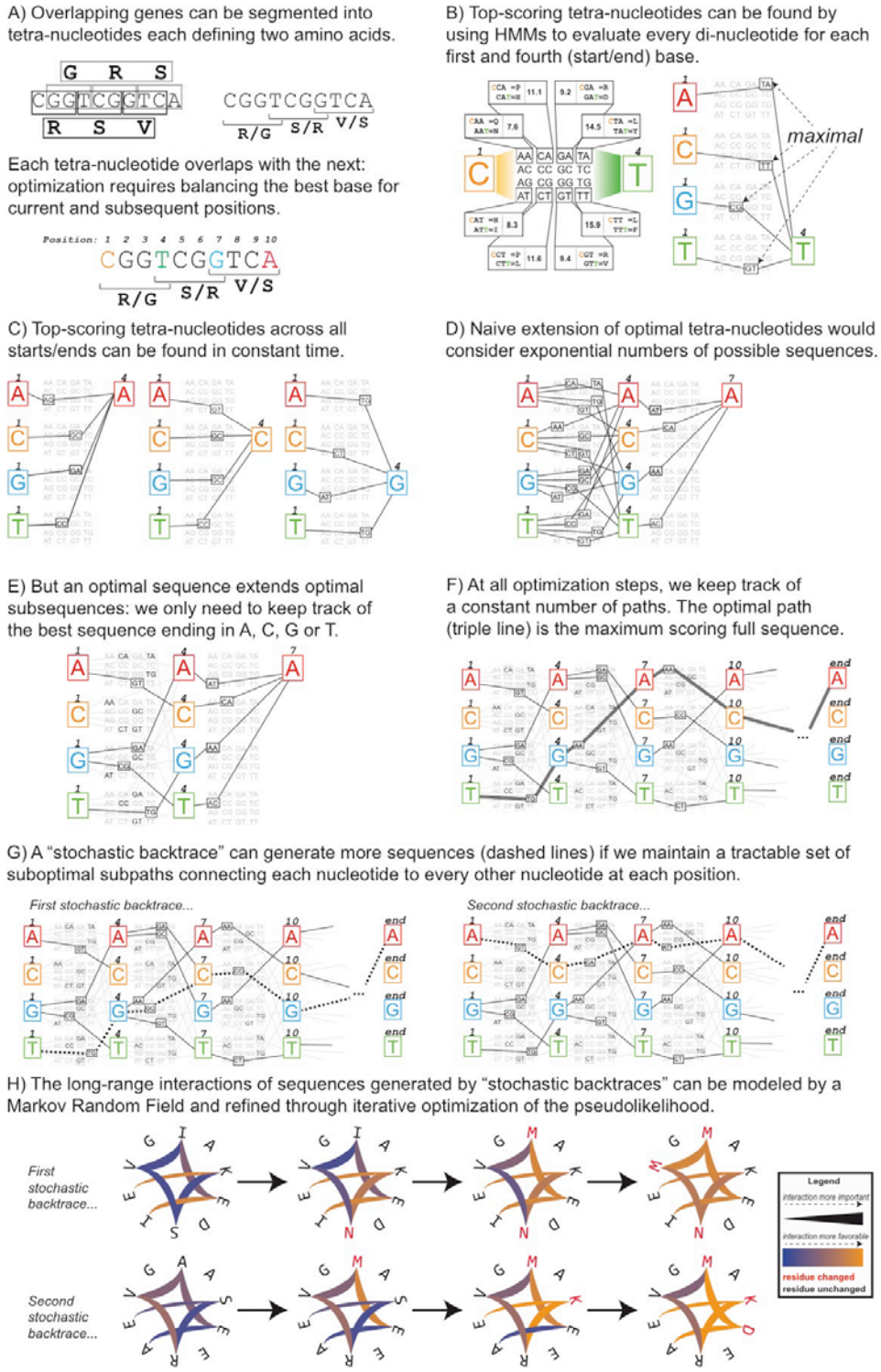

Figure S1. A step-by-step illustration of the CAMEOS approach to designing overlapping sequences. Nucleotides are colored: adenine (A, red), cytosine (C, orange), guanine (G, cyan), thymine (T, green). Steps in the algorithm are described from **(A)** to **(H)**.

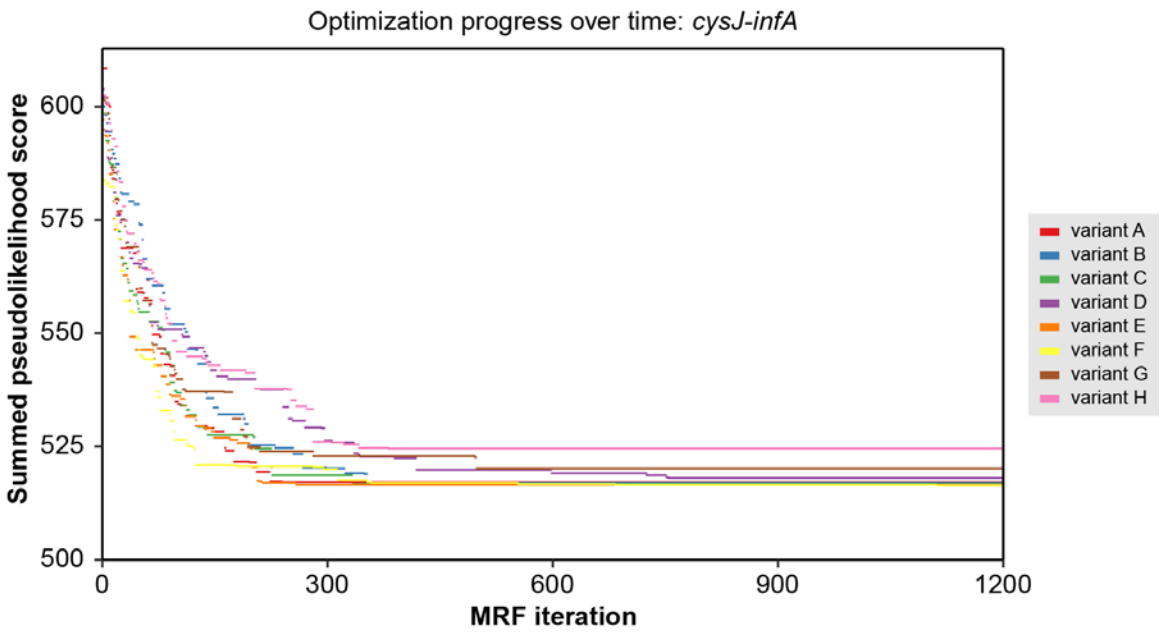

Figure S2. Iterative optimization improves MRF scores. Optimization of the summed MRF pseudolikelihoods of overlapping genes is demonstrated through iterative greedy search for improved long-range interactions during CAMEOS. The sequences are initialized with their optimized HMM values and therefore exhibit high MRF pseudolikelihood values. These values rapidly improve and converge (e.g. after 1000 iterations in this *cysJ-infA* run) by the end of the optimization.

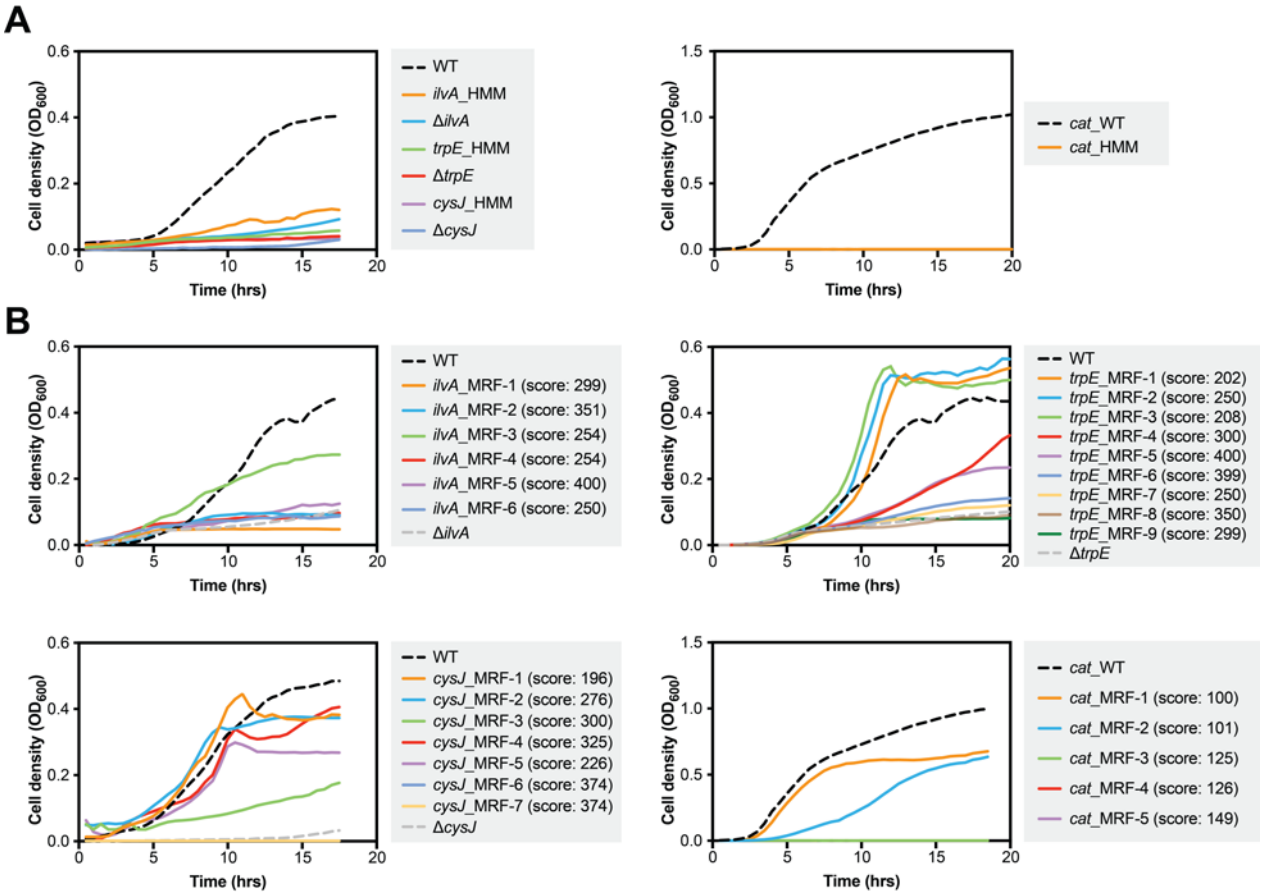

Figure S3. Incorporating MRF models improves the design of functional variants. Growth curves of clones with different designed variants optimized using HMM or MRF models are shown. **(A)** *ilvA*, *trpE*, and *cysJ* HMM-based designs were tested in auxotrophic *E. coli* strains for isoleucine (Δ*ilvA*), tryptophan (Δ*trpE*), and cysteine (Δ*cysJ*), respectively, in minimal media. The *cat* (chloramphenicol acetyltransferase) HMM-based designs were tested in wild-type *E. coli* in LB with standard chloramphenicol selection. **(B)** MRF-based designs for *ilvA*, *trpE*, *cysJ* and *cat* with different MRF pseudolikelihood scores are tested in corresponding strains as in (a) and shown in the four subpanels. Overall, MRF designs show significantly more functional variants with improved activity compared to HMM designs. Data shown are the means of 3-6 independent biological replicate experiments.

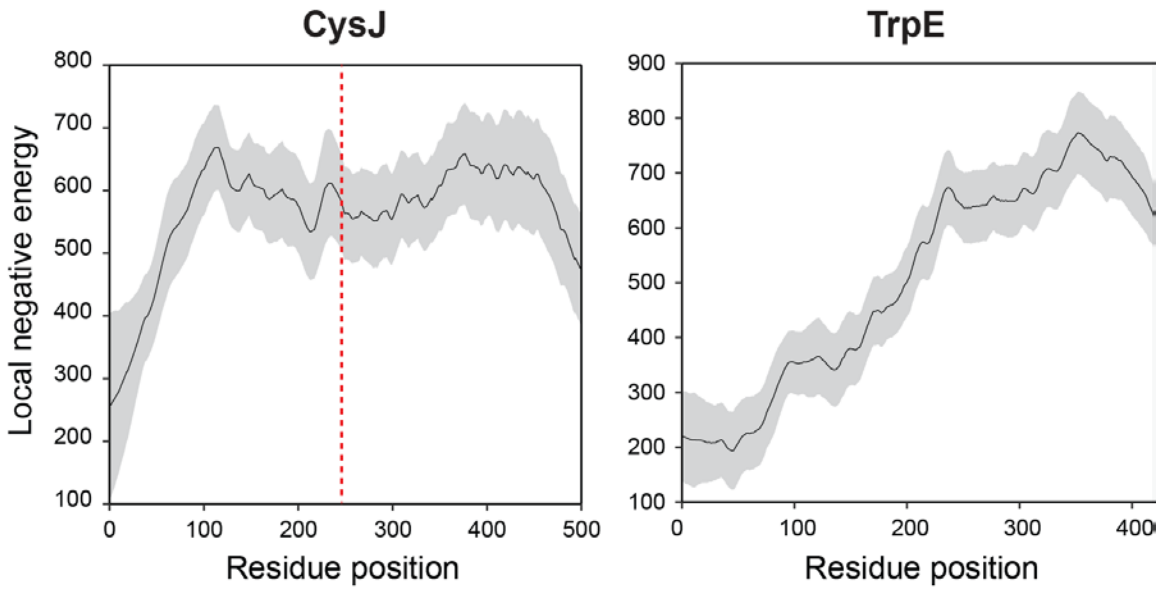

Figure S4. Local energy distribution across designed proteins. The local energies for the CysJ and TrpE proteins are shown on the left and right panels, respectively. The local energy is defined as the sum of the node and edge weights in a Markov Random Field over a 100 residue window for a multiple sequence alignment of the family. The mean is plotted as a black line, with grey regions indicating two standard deviations from the mean. Regions with low values (here most prominently on the N- and C-termini of proteins) likely exhibit low residue-level conservation and/or reduced long-range interactions, suggesting that the region is a promising target for sequence modification and gene overlap designs. The *infA* gene encoding IF1 in our *cysJ-infA* construct is placed starting at residue 246 of CysJ, shown as a red dashed line in the left panel.

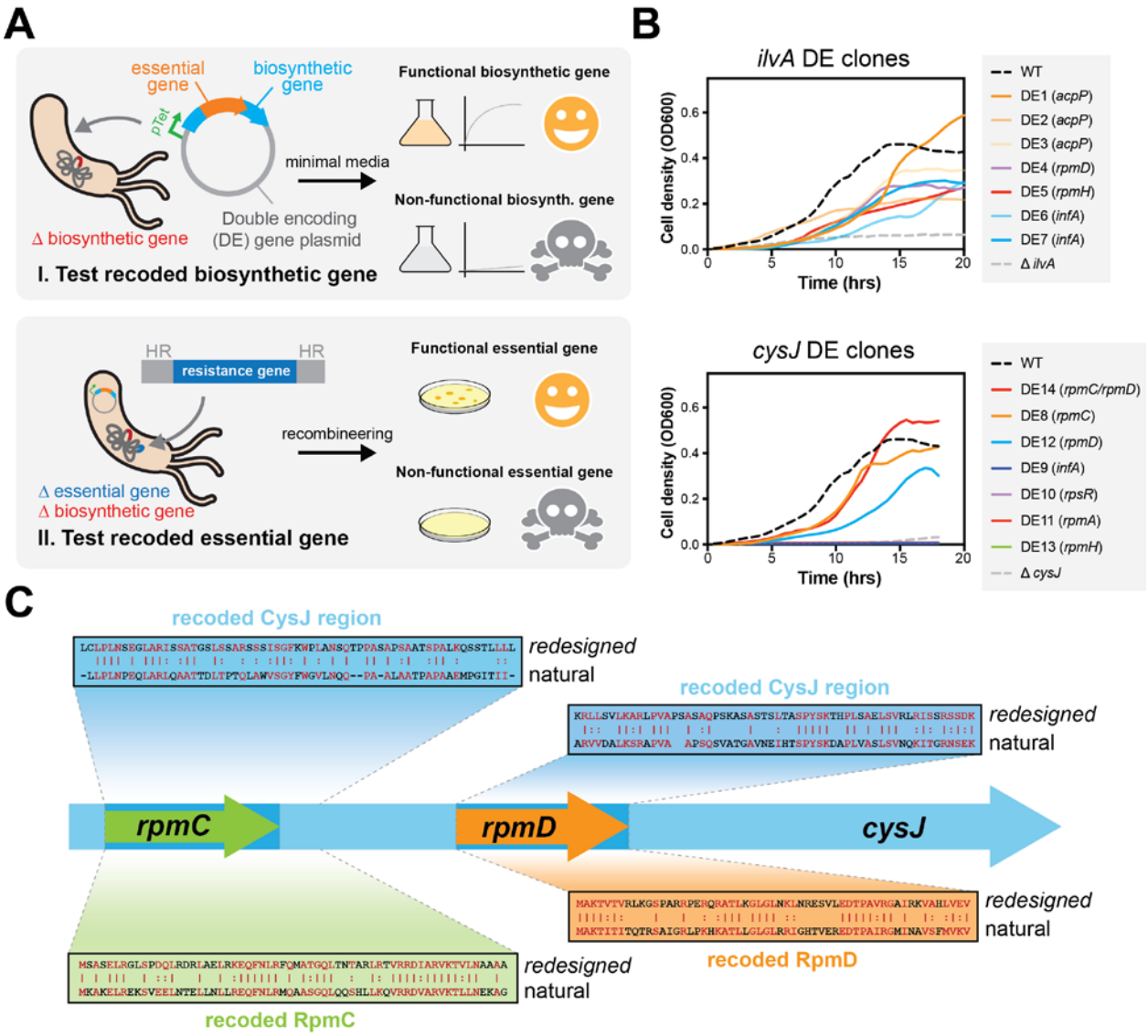

Figure S5. Functional validation of multi-encoding constructs. (A) Experimental strategy for validating double-encoding constructs for biosynthetic genes or essential genes. **(B)** Results for testing the function of co-encoded amino acid biosynthetic genes (*ilvA* and *cysJ*) are shown as growth curves in corresponding auxotrophic strains (Δ*ilvA* and Δ*cysE*) in minimal media. Data are the mean of 3-5 independent biological replicate experiments. **(C)** Sequence layout of clone DE14 showing a recoded *cysJ* with two different essential genes (*rpmC* and *rpmD*) embedded in separate regions of the *cysJ*. The natural (wild-type) and recoded sequences are indicated in the sequence alignments.

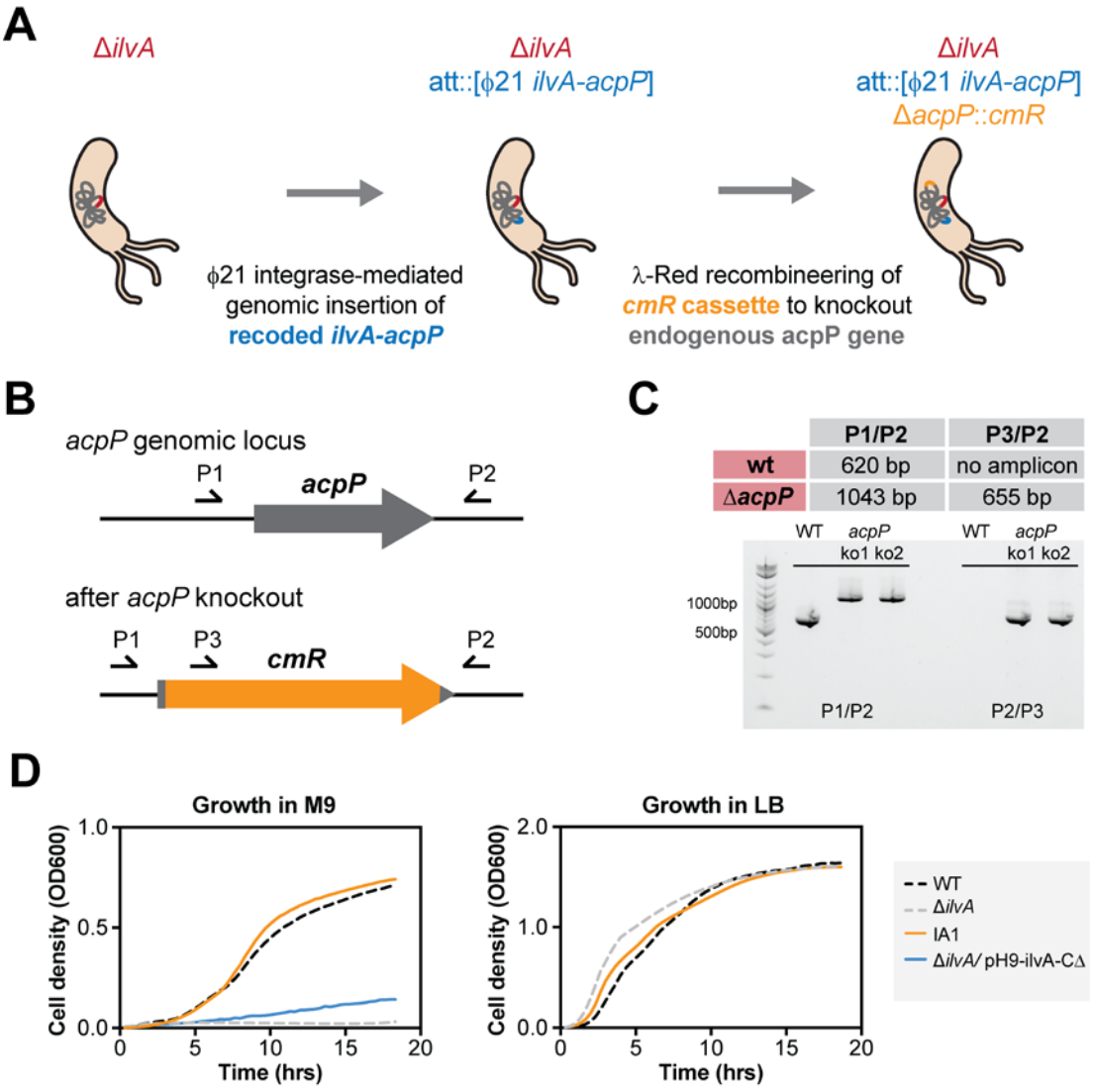

Figure S6. Validation of the IA-1 clone containing a functional and chromosomally integrated *ilvA-acpP* variant. (A) Construction steps used to generate the IA-1 strain to genomically integrate the *ilvA-acpP* cassette and remove the endogenous *acpP* gene. **(B)** Design of PCR primers for verifying genomic *acpP* deletion. **(C)** The table summarizes the expected PCR amplicon size. A gel showing diagnostic PCRs from a wild-type (WT) strain and two isolates (ko1 and ko2) of the *acpP*::*cmR* knockout strain (IA-1). **(D)** Growth of IA-1 compared to control strains and wild-type (WT) cells in M9 minimal media and LB rich media are shown in left and right panels, respectively. The *ΔilvA* contains a genomic knockout of *ilvA*. The pH9-ilvA-CΔ plasmid expresses a C-terminus truncation variant (residues 345-514) of IlvA. The 347-514 region is where the overlapping encoding occurs in the *ilvA-acpP* design. The *ilvA-acpP* variant significantly rescues growth of a *ΔilvA* strain in M9 in contrast to the control strains, thus demonstrating the functional activity of the recoded sequence. Growth curves are mean of 3 to 5 independent experiments.

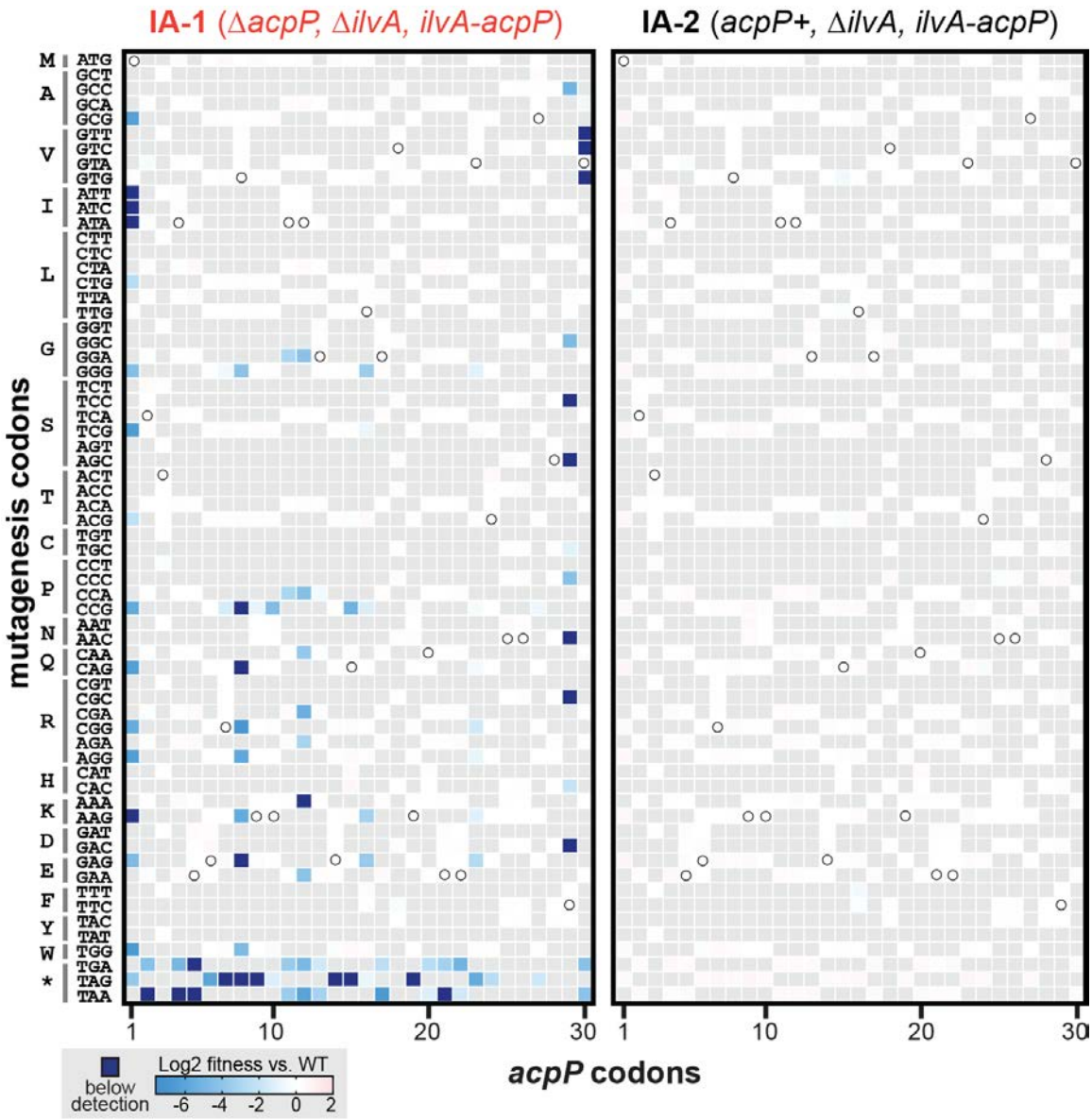

Figure S7. Heatmap of fitness of single-codon mutants from saturation mutagenesis of the *ilvA-acpP* construct in IA-1 and IA-2 strains. X-axis represents codon positions in the *acpP* reading frame. Y-axis represents 64 possible codons grouped by their amino acids. Circles indicate the wild-type *acpP* codons. Since the *ilvA* coding frame was targeted for saturation mutagenesis with NNN-oligos in the *ilvA-acpP* construct, not all *acpP* codons could be generated (shown in grey box). Fitness values are calculated with respect to wild-type sequence. Fitness measurements below the detection limit of deep sequencing are colored dark blue.

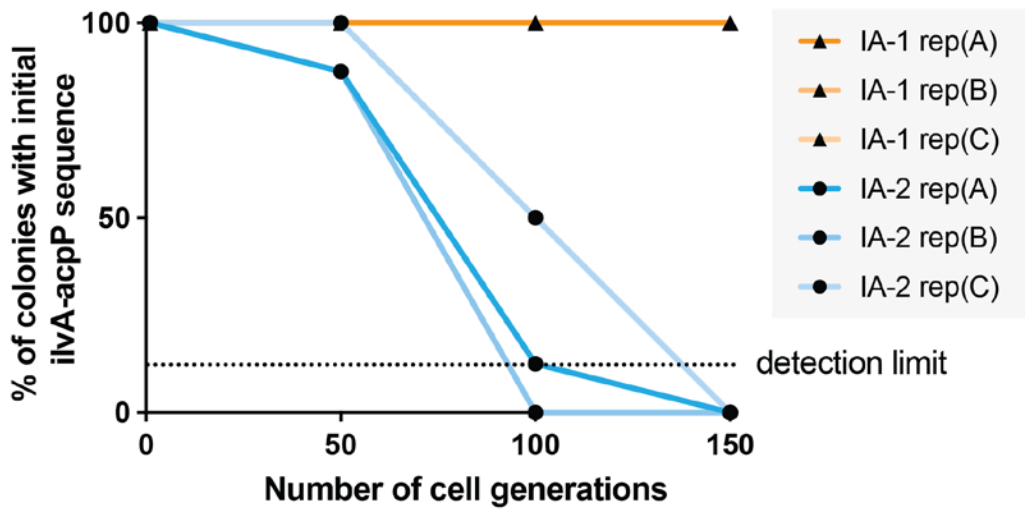

Figure S8. Recoded *ilvA-acpP* resists accumulation of natural mutation under serial growth in laboratory conditions. Three independent replicate cultures (A, B, C) of IA-1 and IA-2 are subject to growth and daily serial dilution over 150 generations. At generations 50, 100, and 150, we randomly isolated 8 colonies from each culture and examined their *ilvA-acpP* region by Sanger sequencing. The y-axis is the percent of colonies that had the initial *ilvA-acpP* sequence.

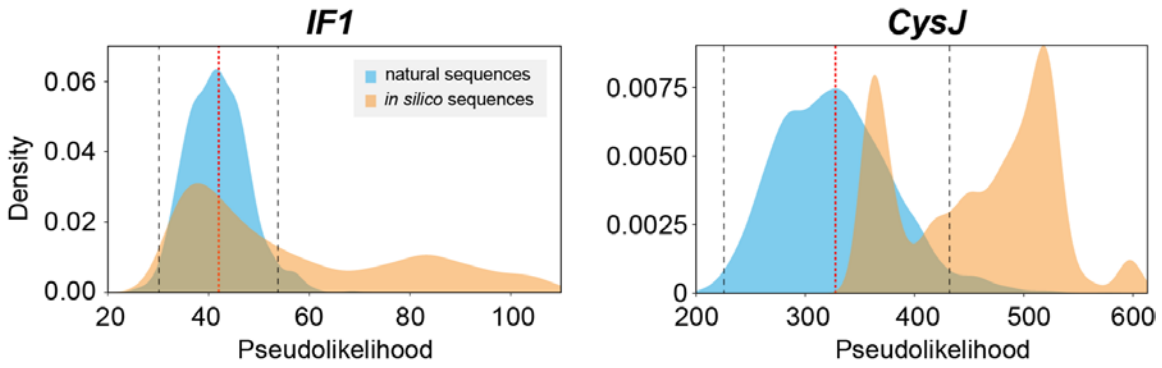

Figure S9. Variants in the *cysJ-infA* library share similar ranges of pseudolikelihood scores with natural IF1 and CysJ sequences. A Gaussian kernel density estimation is used to display the density of pseudolikelihood values of CAMEOS-designed IF1 or CysJ in the *cysJ-infA* library (orange) compared to naturally-occurring sequences in the multiple sequence alignment used to train the Markov Random Field (cyan). The natural sequence distributions are symmetric and appear roughly normally distributed. The mean (red dotted line) and two times the standard deviation from the mean (black dashed lines) are shown for each natural protein variant distribution. Our synthesized variants span a range of pseudolikelihood values (by design) with significant overlap with the naturally-occurring sequences. A small fraction of high pseudolikelihood valued synthetic variants are not shown in the plot to improve visualization.

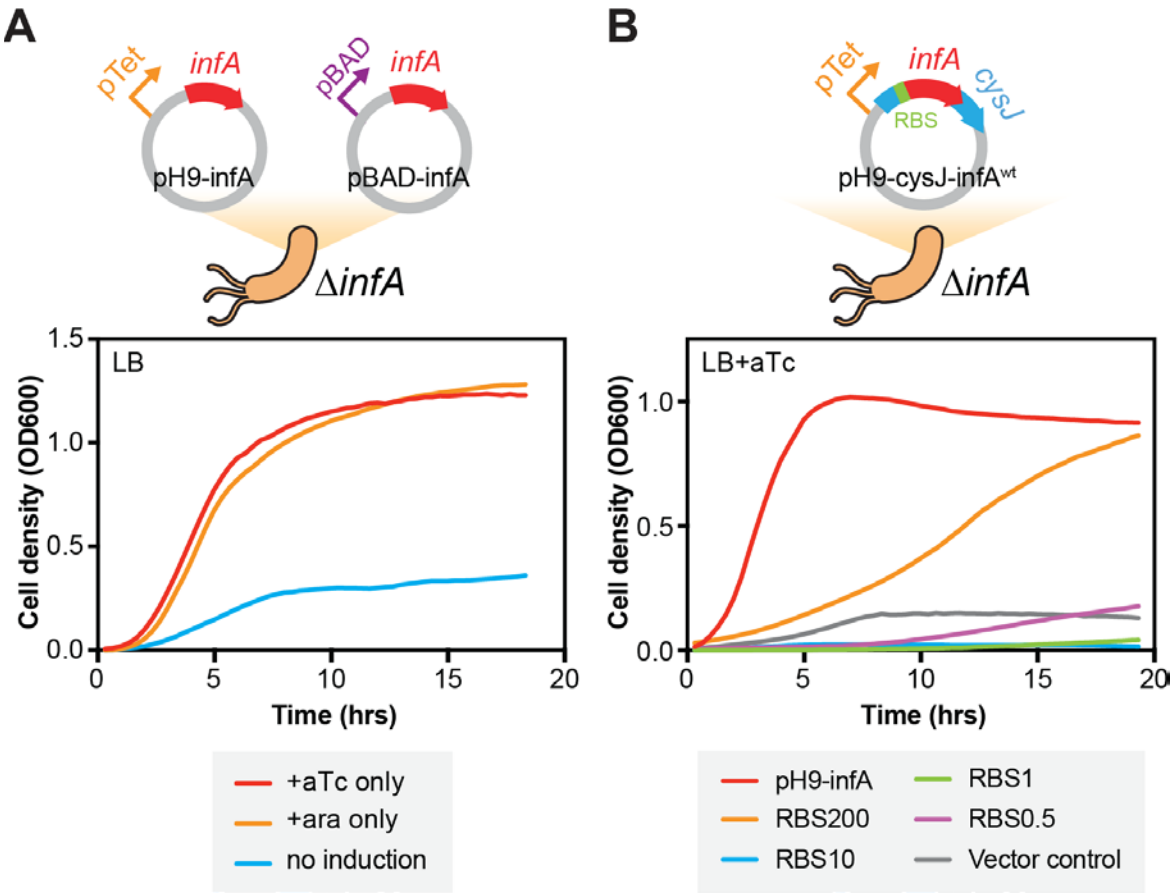

Figure S10. A high-throughput selection strategy to test functional IF1 and optimize internal RBS sequences. (A) A wild-type IF1 (encoded by *infA*) is placed in inducible plasmids pH9 or pBAD. Growth of a Δ*infA* strain can be significantly improved by induction with either arabinose (ara) or anhydrotetracycline (aTc) to express the wild-type *infA*. **(B)** Verifying internal RBS sequences for effective expression of co-encoded *infA*. Embedded wild-type *infA* was constructed with different upstream RBS (still within *cysJ* gene) in the pH9-cysJ-infA^{WT} plasmid. RBS variants are named by their relative *in silico* predicted strength of translation initiation. The RBS200 sequence (corresponding to an output of 200 from the RBS Calculator) showed the best translation of embedded *infA* and was used subsequently in the pH9-cysJ-infA variant library. Each variant is tested for the ability to rescue growth of an Δ*infA* strain. Growth curves are the mean of 3 independent biological replicates.

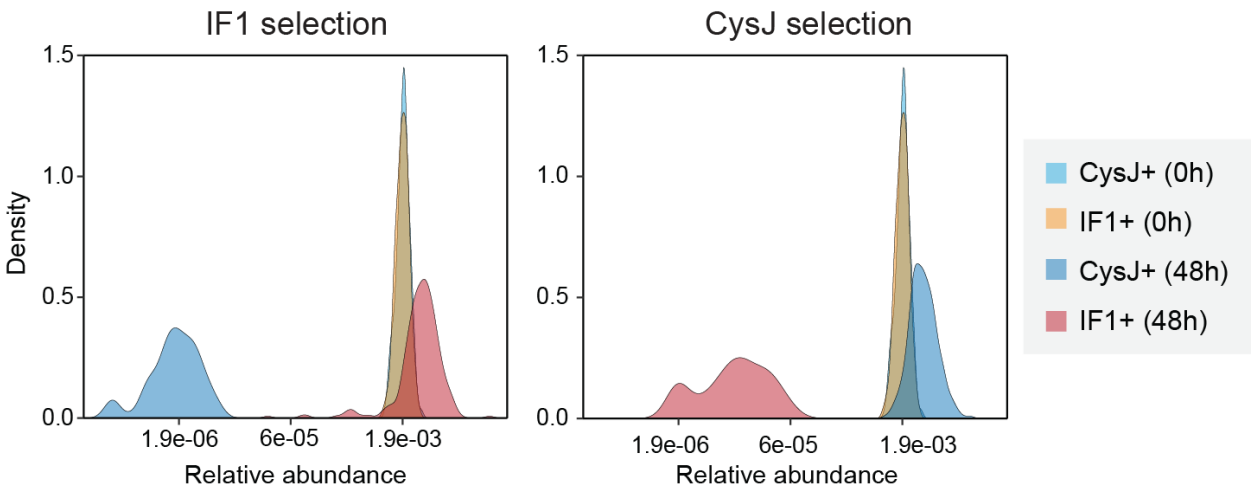

Figure S11. Selection of functional CysJ and IF1 variants. Allele frequencies of positive control clones, CysJ+ and IF1+, were measured before (0h) and after 48-hour selection (48h). At 0h, the relative abundance of CysJ+ and IF1+ are the same. After selection, each population was enriched in their corresponding condition but selected against in the other condition. IF1 selection relies on growth on MOPS plate with cysteine and without arabinose supplement (no IF1 induction) while CysJ selection relies on growth on MOPS plate without cysteine and with arabinose (to induce IF1).

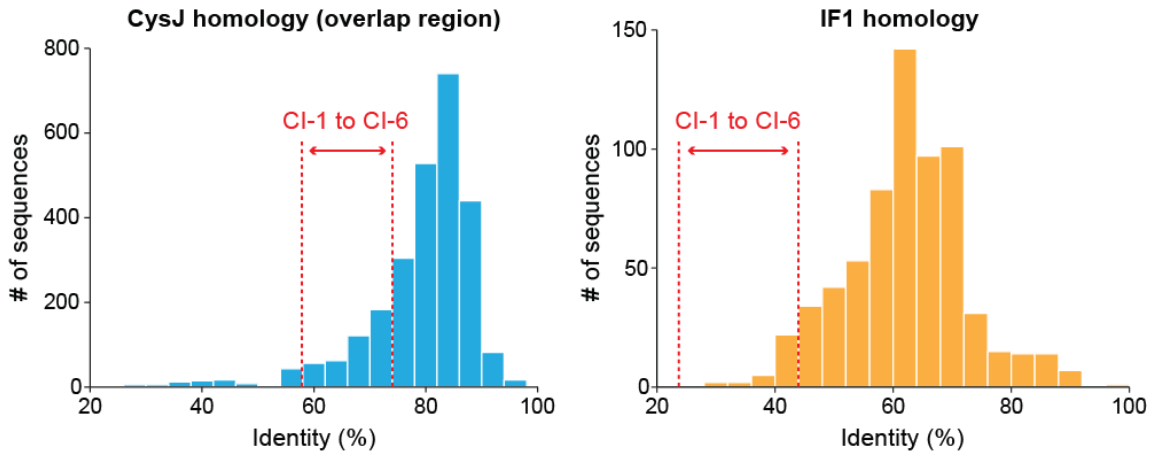

Figure S12. Sequence identities of natural and synthetic CysJ and IF1 proteins. Histograms of sequence identities from naturally occurring CysJ (left, cyan) and IF1 (right, orange) proteins are shown. These protein sequences were taken from the multiple sequence alignment used to train the MRFs, which parameterized the sequence design of the *cysJ-infA* array library. Sequence identity in the case of CysJ is shown only for the region that was recoded. Red dashed lines in both histograms indicate regions of sequence identity exhibited by variants that were found to be functional (CI-1 to CI-6)

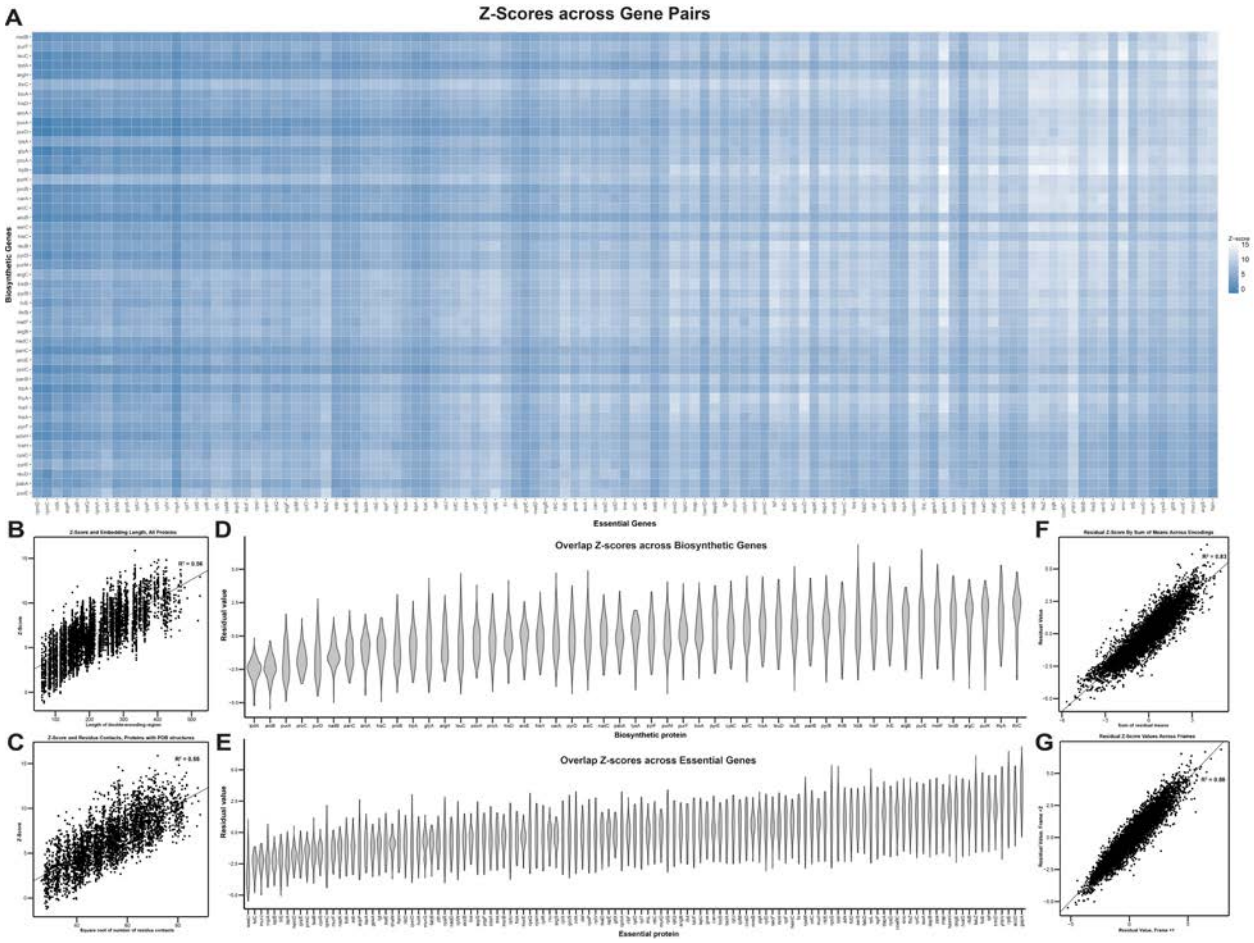

Figure S13. Computational Survey of CAMEOS performance. (A) A heatmap of best-achieved z-scores across pairs of essential (x-axis) and biosynthetic (y-axis) proteins. Proteins on axes are sorted by length. **(B)** Linear regression between the Z-score of a protein pair's CAMEOS embedding and the length of the overlap region. **(C)** Linear regression between the Z-score of a protein pair's CAMEOS embedding and the minimum number of residue contacts in a window of the embedding size across the length of the protein. **(D)** Violin plots of the distributions of residual values from the fit in **(B)** across all essential protein pairs for each biosynthetic protein family. **(E)** Violin plots of the distributions of residual values from the fit in **(B)** across all biosynthetic protein pairs for each essential protein family. **(F)** Linear regression of the residuals from the analysis in **(B)** and the means of residual values in **(D)/(E)** demonstrating high-concordance between the sum of these means and the residual deviation for individual protein pairs. **(G)** Linear regression of the residuals from the analysis in **(B)** across frames, demonstrating high correlation between possible frames.

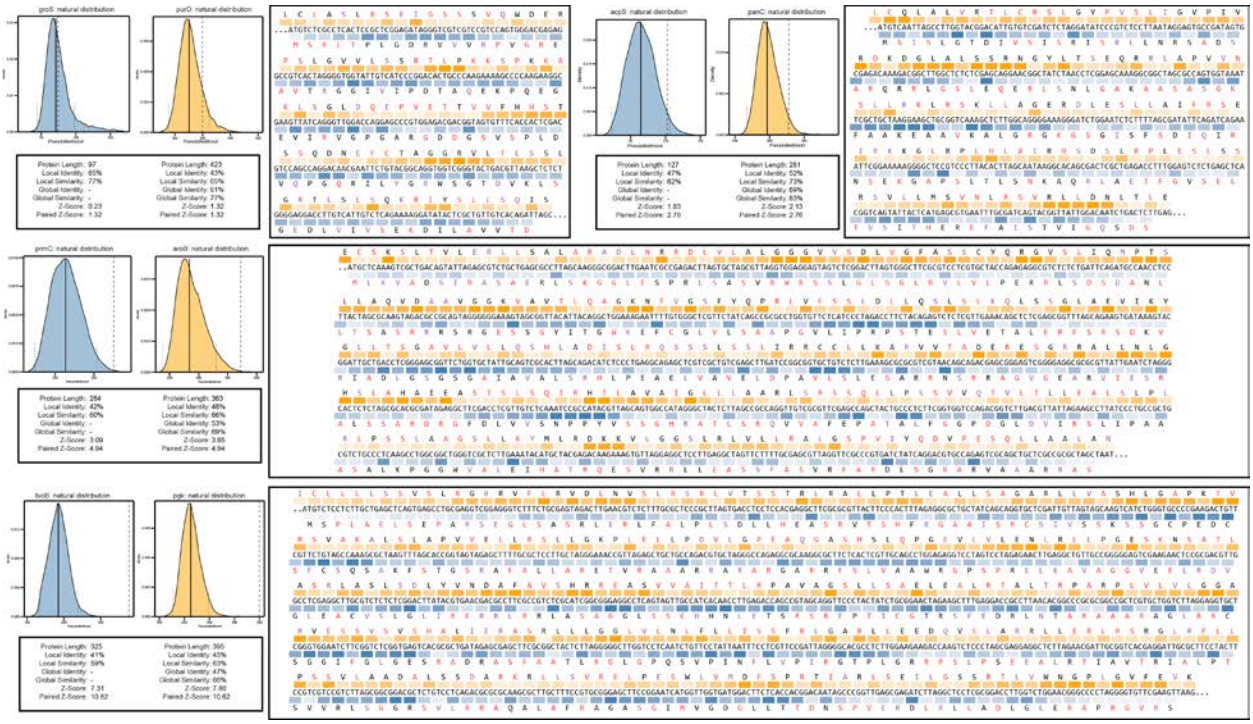

Figure S14. Illustrative examples of double-encoding solutions. Overlap encodings and their translations are shown in both frames for several solutions resulting from our large-scale computational analysis of protein pairs. Selected protein pairs are groS/purD, acpS/panC, prmC/aroB, and bioB/pgk. Histograms demonstrate the natural sequence range of pseudolikelihoods derived from multiple sequence alignments. A single line demarcates the median of the distribution with smaller dashed lines showing 2 MAD deviations from the median. A long dashed line indicates the pseudolikelihood of the considered sequence. Protein translations are demonstrated in two frames: red letters indicate residues that differ from those found in the top BLAST-hit for the embedded portion of the protein. Purple letters indicate similar amino acids at those positions, while black letters are identical. The intensity of the blue/orange coloring in the box at each position reflects the conservation of the amino acids (defined as sequence entropy), with darker shading indicating more conserved residues.

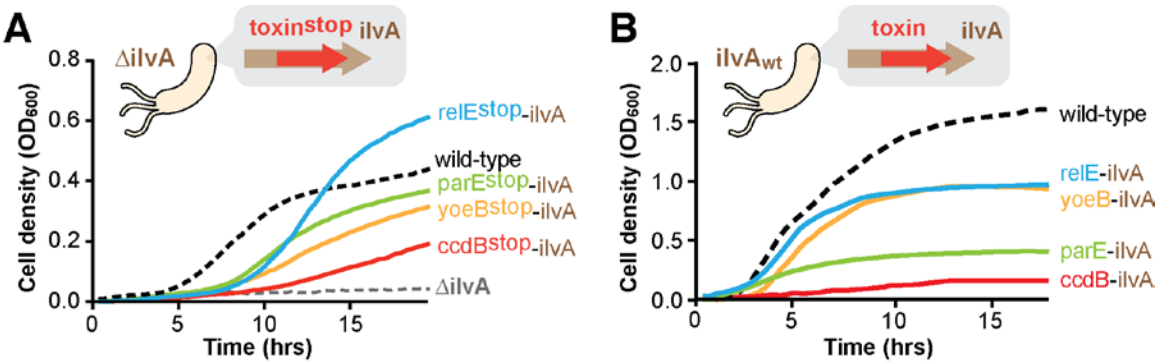

Figure S15. *ilvA*-toxin constructs display varying activity level of *ilvA* and toxins individually. **(A)** Growth curves in M9 minimal media of different *ilvA*-*toxin*^{stop} constructs in a Δ*ilvA* strain are displayed. Different *ilvA* recodings show varying degrees of IlvA function. **(B)** Growth curves in rich media of different *ilvA*-*toxin* constructs in a wildtype strain are displayed. Different encoded toxins show varying degrees of growth inhibition. Data shown are the means of 3 independent biological replicate experiments.

References

1. B. G. Barrell, G. M. Air, C. A. Hutchison 3rd, Overlapping genes in bacteriophage φX174. *Nature* **264**, 34–41 (1976). doi:10.1038/264034a0 Medline

2. C. A. Spencer, R. D. Gietz, R. B. Hodgetts, Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature* **322**, 279–281 (1986). doi:10.1038/322279a0 Medline

3. I. B. Rogozin, A. N. Spiridonov, A. V. Sorokin, Y. I. Wolf, I. K. Jordan, R. L. Tatusov, E. V. Koonin, Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* **18**, 228–232 (2002). doi:10.1016/S0168-9525(02)02649-5 Medline

4. C. R. Sanna, W.-H. Li, L. Zhang, Overlapping genes in the human and mouse genomes. *BMC Genomics* **9**, 169 (2008). doi:10.1186/1471-2164-9-169 Medline

5. Z. I. Johnson, S. W. Chisholm, Properties of overlapping genes are conserved across microbial genomes. *Genome Res.* **14**, 2268–2272 (2004). doi:10.1101/gr.2433104 Medline

6. M. Mizokami, E. Orito, K. Ohba, K. Ikeo, J. Y. N. Lau, T. Gojobori, Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.* **44** (suppl. 1), S83–S90 (1997). doi:10.1007/PL00000061 Medline

7. J. Choi, Z. Xu, J. H. Ou, Triple decoding of hepatitis C virus RNA by programmed translational frameshifting. *Mol. Cell. Biol.* **23**, 1489–1497 (2003). doi:10.1128/MCB.23.5.1489-1497.2003 Medline

8. T. Miyata, T. Yasunaga, Evolution of overlapping genes. *Nature* **272**, 532–535 (1978). doi:10.1038/272532a0 Medline

9. E. Simon-Loriere, E. C. Holmes, I. Pagán, The effect of gene overlapping on the rate of RNA virus evolution. *Mol. Biol. Evol.* **30**, 1916–1928 (2013). doi:10.1093/molbev/mst094 Medline

10. B. Csörgo, T. Fehér, E. Tímár, F. R. Blattner, G. Pósfai, Low-mutation-rate, reduced-genome Escherichia coli: An improved host for faithful maintenance of engineered genetic constructs. *Microb. Cell Fact.* **11**, 11 (2012). doi:10.1186/1475-2859-11-11 Medline

11. F. K. Balagaddé, L. You, C. L. Hansen, F. H. Arnold, S. R. Quake, Long-term monitoring of bacteria undergoing programmed population control in a microchemostat. *Science* **309**, 137–140 (2005). doi:10.1126/science.1109173 Medline

12. B. R. Jack, S. P. Leonard, D. M. Mishler, B. A. Renda, D. Leon, G. A. Suárez, J. E. Barrick, Predicting the Genetic Stability of Engineered DNA Sequences with the EFM Calculator. *ACS Synth. Biol.* **4**, 939–943 (2015). doi:10.1021/acssynbio.5b00068 Medline

13. A. Chavez, B. W. Pruitt, M. Tuttle, R. S. Shapiro, R. J. Cecchi, J. Winston, B. M. Turczyk, M. Tung, J. J. Collins, G. M. Church, Precise Cas9 targeting enables genomic mutation

prevention. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3669–3673 (2018). doi:10.1073/pnas.1718148115 Medline

14. J. W. Lee, C. T. Y. Chan, S. Slomovic, J. J. Collins, Next-generation biocontainment systems for engineered organisms. *Nat. Chem. Biol.* **14**, 530–537 (2018). doi:10.1038/s41589-018-0056-x Medline

15. See supplementary materials.

16. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011). doi:10.1002/prot.22934 Medline

17. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013). doi:10.1073/pnas.1314045110 Medline

18. H. M. Salis, The ribosome binding site calculator. *Methods Enzymol.* **498**, 19–42 (2011). doi:10.1016/B978-0-12-385120-8.00002-4

19. C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, D. Karlin, Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J. Virol.* **83**, 10719–10736 (2009). doi:10.1128/JVI.00595-09 Medline

20. E. D. Kelsic, H. Chung, N. Cohen, J. Park, H. H. Wang, R. Kishony, RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Syst.* **3**, 563–571.e6 (2016). doi:10.1016/j.cels.2016.11.004 Medline

21. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015). doi:10.1038/nprot.2015.053 Medline

22. V. Opuu, M. Silvert, T. Simonson, Computational design of fully overlapping coding schemes for protein pairs and triplets. *Sci. Rep.* **7**, 15873 (2017). doi:10.1038/s41598-017-16221-8 Medline

23. V. Kolmogorov, Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1568–1583 (2006). doi:10.1109/TPAMI.2006.200 Medline

24. 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**, 1364–1368 (2003). doi:10.1126/science.1089427 Medline

25. C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018). doi:10.1126/science.aao5167 Medline

26. F. St-Pierre, L. Cui, D. G. Priest, D. Endy, I. B. Dodd, K. E. Shearwin, One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013). doi:10.1021/sb400021j Medline

27. Á. Nyerges, B. Csörgő, I. Nagy, B. Bálint, P. Bihari, V. Lázár, G. Apjok, K. Umenhoffer, B. Bogos, G. Pósfai, C. Pál, A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2502–2507 (2016). doi:10.1073/pnas.1520040113 Medline

28. S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, C. Yeats, InterPro: The integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009). doi:10.1093/nar/gkn785 Medline

29. S. Deorowicz, A. Debudaj-Grabysz, A. Gudyś, FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964 (2016). doi:10.1038/srep33964 Medline

30. P. Jehl, F. Sievers, D. G. Higgins, OD-seq: Outlier detection in multiple sequence alignments. *BMC Bioinformatics* **16**, 269 (2015). doi:10.1186/s12859-015-0702-1 Medline

31. S. Seemayer, M. Gruber, J. Söding, CCMpred—Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130 (2014). doi:10.1093/bioinformatics/btu500 Medline