Pervasive fold switching in a ubiquitous protein superfamily

Authors: Lauren L. Porter[1,2,*], Allen K. Kim[1], Loren L. Looger[3], Anaya Majumdar[4], and Mary Starich[2]

[1]National Library of Medicine, National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

[2]National Heart, Lung, and Blood Institute, Biochemistry and Biophysics Center, National Institutes of Health, Bethesda, MD 20892, USA

[3]Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA 20147, USA

[4]The Johns Hopkins University Biomolecular NMR Center, The Johns Hopkins University, Baltimore, MD 21218, USA

[*]Corresponding author: lauren.porter@nih.gov

Abstract: Fold-switching proteins challenge the one-sequence-one-structure paradigm by adopting multiple stable folds. Nevertheless, it is uncertain whether fold switchers are naturally pervasive or rare exceptions to the well-established rule. To address this question, we developed a predictive method and applied it to the NusG superfamily of >15,000 transcription factors. We estimate that a substantial population (25%) of the proteins in this family switch folds. Circular dichroism and nuclear magnetic resonance spectroscopies of 10 sequence-diverse variants confirmed our predictions. Subsequently, we leveraged family-wide predictions to determine both conserved contacts and taxonomic distributions of fold-switching proteins. Our results indicate that fold switching is pervasive in the NusG superfamily and that the one-sequence-one-structure protein folding paradigm significantly biases protein structure prediction strategies.

Main Text: Fold-switching proteins remodel their secondary and tertiary structures and change their functions in response to cellular stimuli (*1*). Their large, sometimes reversible, conformational changes challenge the long-held paradigm that globular proteins have a single fold specified by their amino acid sequences (*2*). Fold switchers regulate diverse biological processes (*3*) and are associated with human diseases such as cancer (*4*), malaria (*5*), and COVID-19 (*6*).

Whereas computational approaches such as coevolutionary analysis (*7*) and deep learning (*8*), have greatly improved rapid prediction of secondary and tertiary structure for single-fold proteins (*9*), similar methods to classify fold switchers have lagged. This comparative lack of progress arises primarily from the small number of experimentally observed fold switchers (<100), hampering the discovery of generalizable characteristics that distinguish fold switchers from single folders. As a result, essentially all naturally occurring fold switchers have been discovered by chance (*10*), leaving their natural abundance an open question.

To assess the natural abundance, evolutionary conservation, and taxonomic diversity of fold-switching proteins, we sought to predict them from their sequences. Previous work has shown that discrepancies between homology-based secondary structure predictions often indicate protein fold switching (*11-13*). We leveraged these discrepancies to discover disparate fold-switching proteins in the NusG protein superfamily, the only family of transcriptional regulators known to be conserved from bacteria to humans (*14*). Housekeeping NusGs (hereafter called NusGs) exist in nearly every known bacterial genome and associate with transcribing RNA polymerase (RNAP) at essentially every operon, where they promote transcription elongation. By contrast, specialized NusGs ($NusG^{SP}$s), such as UpxY, LoaP, and RfaH, promote transcription elongation at specific operons only (*15*). Atomic-level structures of RfaH and NusG have been determined (**Fig. 1A**). They share a two-domain architecture with an N-terminal NGN domain that binds RNAP, and a C-terminal β-roll domain. By contrast, the C-terminal domain (CTD) of *Escherichia coli* RfaH switches between two disparate folds: an α-helical hairpin that inhibits RNAP binding except at operon polarity suppressor (*ops*) DNA sites and a β-roll that binds the S10 ribosomal subunit, fostering efficient translation (*16*). This reversible change in structure and function is triggered by binding to both *ops* DNA and RNAP (*17*).

We observed that JPred4 (*18*) secondary structure predictions discriminate between the sequences of RfaHs and NusGs with experimentally determined atomic-level structures (**Figs. 1A&B**). These sequence-based calculations consistently indicate that NusG CTD sequences fold into β-strands connected by coils, whereas the *E. coli* RfaH CTD assumes a mixture of α-helix, β-strand, and coil. Thus, our results suggest that JPred4 can distinguish between the sequences of single-folding NusGs and the fold-switching $NusG^{SP}$, RfaH.

To test the generality of our secondary-structure-based approach, we collected 15,516 non-redundant NusG/$NusG^{SP}$ sequences from an iterative BLAST (*19*) search (**Methods**) and tested our predictive method on each hit from the search. In total, 25% of proteins in the NusG superfamily were predicted to switch folds (**Figures 1C and S1, Data S1**), a considerable subpopulation with over 3500 sequences.

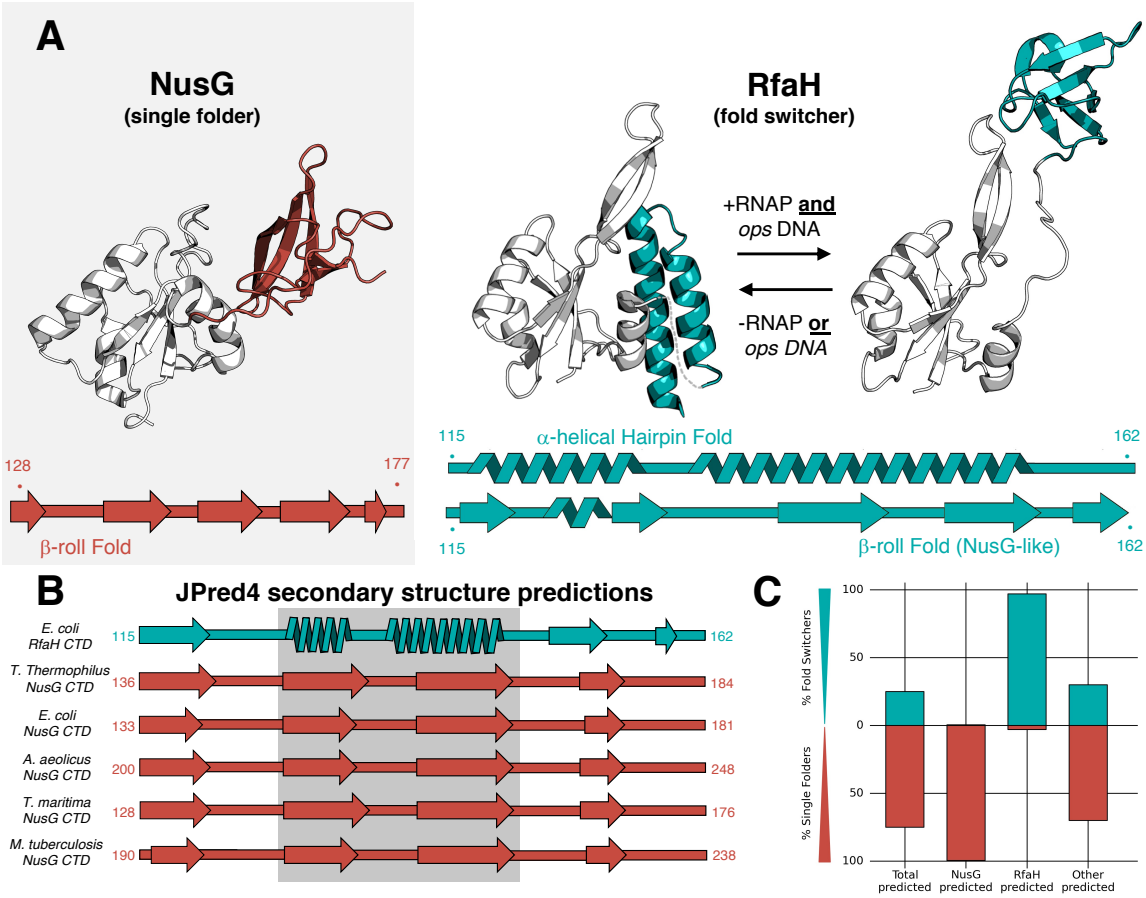


Fig. 1. Sequence-based secondary structure predictions discriminate between fold switchers and single folders. (A). Experimentally determined folds and secondary structures of single folder NusG and the autoinhibited/active NusG^{SP}, RfaH (α-helical hairpin/β-roll folds, respectively). Dotted line represents the NTD-CTD linker missing in the RfaH crystal structure. NusG/RfaH CTDs are colored teal/red; NTDs are gray. **(B).** JPred4 secondary structure predictions discriminate between RfaH and NusGs with experimentally determined structures (teal and red, respectively). Gray box highlights secondary structure prediction discrepancies. Residue numbers are at each end of the secondary structure diagrams. **(C).** Of the sequences in the NusG family, 25% were predicted to switch folds. As expected, nearly all (99.5%) genomically verified housekeeping NusG sequences within our dataset were predicted to be single folders. Furthermore, 97% of RfaH-like sequences identified previously (15) and 30% of the remaining sequences with high-confidence predictions (Other) were predicted to switch folds.

To determine the false-negative and false-positive rates of our predictions, we exploited known operon structures of NusG and several specialized orthologs (*15*) as an orthogonal method to annotate sequences as NusGs or NusG^{SP}s. We mapped the sequences used for prediction to sequenced bacterial genomes (Ensembl; **Methods**) and analyzed each sequence's local genomic environment for signatures of co-regulated genes. Of our 15,516 total sequences, 5,435 mapped to Ensembl contexts consistent with housekeeping NusG function. Only 30 of these were predicted to switch folds (**Figure 1C**), suggesting a false-positive rate of 0.6% for fold-switch predictions. Performing a similar calculation in previously identified RfaHs (*15*), we found that of the 1,078 in **Data S1**, 31 were predicted not to switch folds (**Figure 1C**). These results suggest that fold switching is widely conserved among RfaHs, which, if correct, indicates a false positive rate of 3% (31/1078). Of the remaining sequences with high-confidence predictions (**Methods**), 30% were predicted to switch folds (**Figure 1C**).

A representative group of variants with dissimilar sequences was then selected for experimental validation. First, all NusG-superfamily sequences were clustered and plotted on a force-directed graph, hereafter called NusG sequence space (**Figure 2A, Data S1, Figure S2**). The map of this space revealed that some putative fold-switching/single-folding nodes cluster together within sequence space (upper/lower groups of interconnected nodes), while other regions had mixed predictions (left/right groups of interconnected nodes). Candidates selected for experimental validation came from distinct nodes, had diverse genomic annotations, and originated from different bacterial phyla (**Table S1, Figure S3**).

Circular dichroism (CD) spectra of 10 full-length variants were collected. We expected the spectra of fold switchers to have more helical content than single folders because their CTDs have completely different structures (RfaH: all α-helix, NusG: all β-sheet), while the secondary structure compositions of their single-fold NTDs are expected to be essentially identical. *E. coli* RfaH (variant #3) and *E. coli* NusG (variant #9) were initially compared because their atomic-level structures have been determined previously (*20, 21*). As expected, their CD spectra were quite different (**Figure S4**): *E. coli* RfaH had a substantially higher α-helix:β-strand ratio (1.1, **Figure 2B, variant #3**) than *E. coli* NusG (0.54, **Figure 2B, variant #9**) – consistent with solved structures.

All 10 of our predictions were consistent with their corresponding CD spectra (**Figure 2B, Table S1**). Specifically, in addition to *E. coli* RfaH, five other predicted fold switchers had RfaH-like CD spectra: two RfaHs (variants #2, #6), a LoaP (variant #1), an annotated NusG (variant #4), and an annotated "NGN domain-containing protein" (variant #5). Furthermore, the remaining three predicted single folders had NusG-like CD spectra: two annotated NusGs (variants #8, #10) and one UpbY/UpxY (variant #7).

We then assessed whether putative fold-switching CTDs could assume β-sheet folds in addition to the α-helical conformations suggested by CD. Previous work (*16*) has shown that the full-length RfaH CTD folds into an α-helical hairpin while its isolated CTD folds into a stable β-roll. Thus, we determined the CD spectra of five isolated CTDs: three from putative fold switchers and two from putative single folders. All of them had low helical content and high β-sheet content (**Figure S5**), strongly suggesting that the CTDs of all three predicted fold switchers can assume both α-helical hairpin to β-roll topologies.

Two variants were then characterized at higher resolution using nuclear magnetic resonance (NMR) spectroscopy. Previous work (*16*) has shown that the isolated CTD of RfaH has a significantly different ^{1}H-^{15}N HSQC than full-length RfaH, whose CTD folds into an α-helical hairpin. Thus, we conducted similar experiments on one single-fold variant (variant #8) and one putative fold switcher (variant #5). We found that the backbone amide resonances of the full-length and CTD forms of variant #8 were nearly superimposable, whereas the full-length and CTD forms of variant #5 shared only 7/58 common backbone amide peaks (**Figure 2C**). This result demonstrates that, as predicted, variant #8 does not switch folds. It is also consistent with the prediction that variant #5 switches folds because large backbone amide shifts indicate either refolding or a large change in protein interface. Both occur in fold-switching RfaH but not in single-folding NusG. Subsequently, we used assigned backbone amide resonances to characterize

the secondary structures of both CTD variants at higher resolution (**Figure S6, Table S2**). Both were consistent with the β-roll fold, again indicating that the CTD of variant #5 may switch folds.

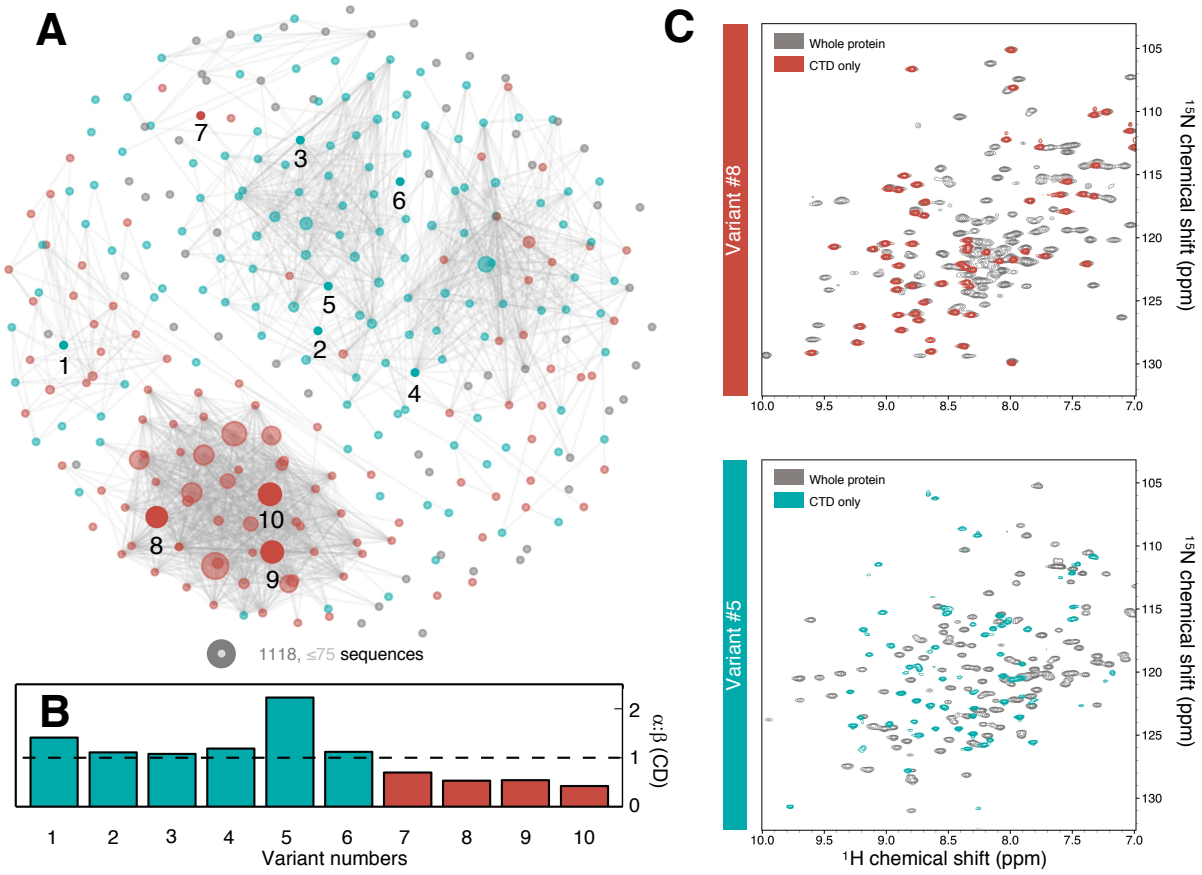


Fig. 2. RfaH/NusG sequence space. (A). Force-directed graph of 15,516 full-length RfaH/NusG sequences. The largest node contains 1118 sequences; all nodes with 75 sequences or fewer are the same (smallest) size. Edges connecting the graph represent an average aligned identity between the sequences in two nodes that exceeds 22%. Nodes labeled in teal/red were predicted to be fold switchers/single folders, on average; gray nodes contained only sequences with low-confidence JPred4 predictions. **(B).** Fraction of α-helix:β-sheet measured from CD. Dotted line (1.0) represents the minimum α-helix:β-sheet ratio for fold switchers. All ratios for predicted fold switchers are above the cutoff; all ratios for predicted single folders fall below. Numerical labels shown in **(A)** correspond to variant numbers. **(C).** The ^{1}H-^{15}N HSQCs of full-length and CTD variants of a putative single-folder (Variant #8) are nearly superimposable, while the HSQCs of full-length and CTD variants of a putative fold switcher (Variant #5) differ significantly.

These results, though a very small proportion of the sequences in this superfamily, support the accuracy of our predictions and demonstrate that:

(1) Some but not all NusG^{SP}s besides RfaH probably switch folds. Specifically, full-length LoaP (variant #1), which regulates the expression of antibiotic gene clusters (*22*), had an RfaH-like CD spectrum, whereas full-length UpbY from *B. fragilis* (variant #7) appears to assume a NusG-like fold.

(2) Some annotated NusGs have RfaH-like CD spectra (variant #4), the result of incorrect annotation. Indeed, the genomic environment of variant #4 (**Methods**) suggests that is a UpxY, not a NusG.

(3) The fold-switching mechanism is conserved among annotated RfaHs with low sequence identity (≤32%, variants #2, #3, and #6), a possibility proposed previously

(*23*), though without experimental validation. Also, "NGN domain-containing protein" variant #5 is genomically inconsistent with being a NusG and is likely another RfaH.

To benchmark the performance of our secondary-structure-based method, we assessed whether coevolutionary and template-based methods could also distinguish between fold switchers and single folders in the NusG superfamily. Specifically, we tested Robetta (*24*), EVCouplings (*25*), and Phyre2 (*26*) on variants #1-6 (**Figure 3A**). All methods predicted only β-strand conformations (**Figure S7**). These predictions included *E. coli* RfaH (variant #3), which assumes an α-helical hairpin in its experimentally determined structure (*20*). The multiple sequence alignments used to generate these predictions contained mixtures of RfaH and NusG sequences, and the resulting residue-residue couplings from Robetta and EVCouplings corresponded with the NusG-like β-roll fold (**Figure 3A**). These results suggest that β-roll couplings present in both single-folding and fold-switching sequences might overwhelm any α-helical couplings unique to fold-switching sequences.

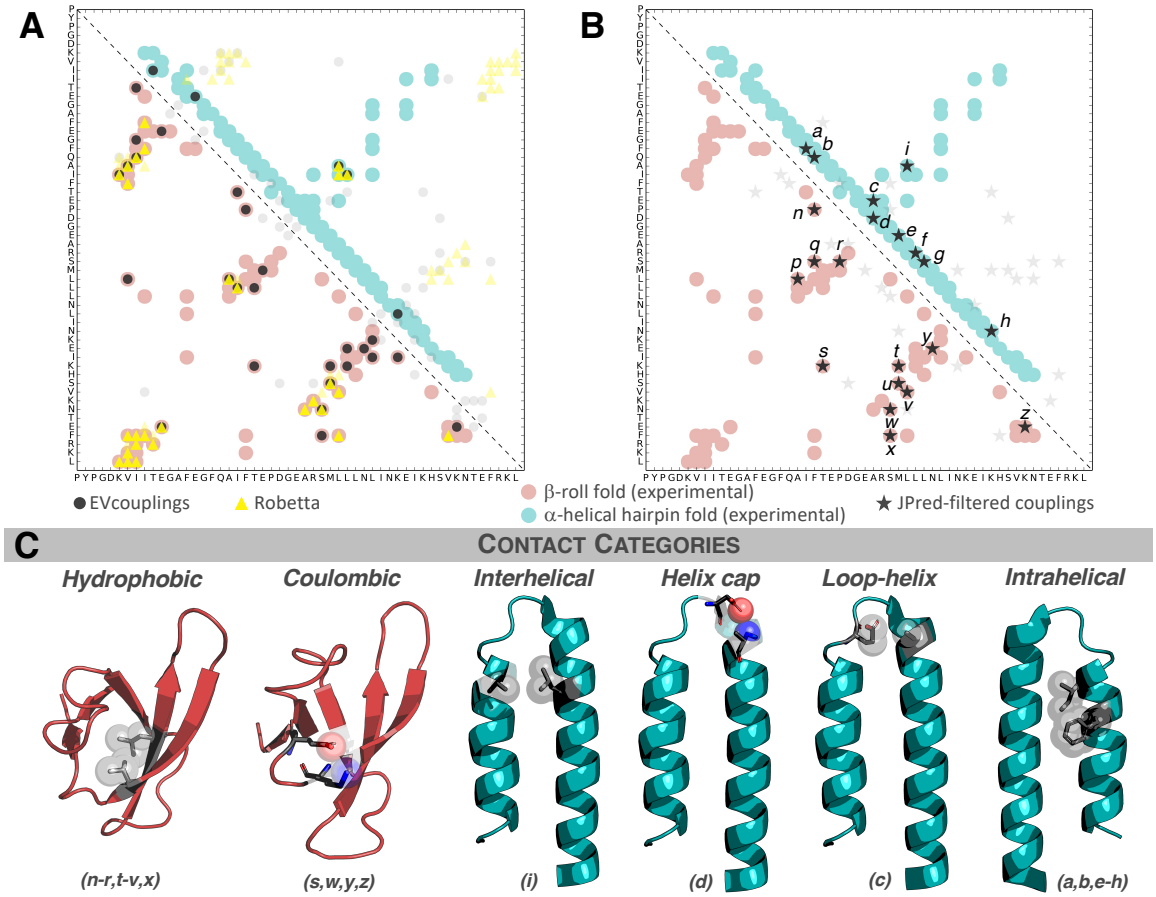


Fig. 3. The single-fold paradigm biases protein structure predictions. (A) Residue couplings for *E. coli* RfaH predicted by EVCouplings (gray circles) and Robetta (yellow triangles) are biased toward the β-roll fold (PDB ID: 2LCL, red circles); no couplings unique to the experimentally determined α-helical hairpin (PDB ID: 2OUG, chain A, teal circles) were identified. Both sets of predictions were calculated from alignments with mixtures of single-folding and fold-switching sequences. Couplings that do not correspond to experimentally observed contacts are lighter. **(B)** Sequences with JPred4 predictions similar to *E. coli* RfaH yielded residue-residue couplings from both the β-roll and the α-helical hairpin folds (gray stars). Italicized letters correspond to individual couplings in the two folds. **(C)** Categories of residue-residue contacts from both folds using the alphabetically labeled contacts in **(B)**, listed below each category.

We then performed coevolutionary analysis on a set of sequences that our method predicted to switch folds. Specifically, we clustered putative fold switchers by their JPred4 predictions and did coevolutionary analysis on the cluster containing the *E. coli* RfaH sequence using GREMLIN (*27*). The residue-residue couplings generated from these sequences differed substantially from the NusG-like couplings generated before (**Figures 3A&B**). Furthermore, GREMLIN couplings calculated from the alignments used by EVCouplings and Robetta corresponded with the β-roll fold only (**Figure S8**), demonstrating that the JPred-filtered sequence alignment—not the GREMLIN algorithm—was responsible for the discovery of alternative contacts**.**

The coevolutionary analysis of putative fold switchers suggests that their CTDs encode contacts unique to both the α-helical and β-roll folds – only one contact was shared between both (*i/q*, **Figures 3B&C**). Several categories of contacts were identified. For the β-roll fold, eight hydrophobic and four Coulombic contacts were observed. Using residue pairs from all JPred4-filtered sequences, we found that 97%/96% of residue pairs making strand contacts *q* and *v* were hydrophobic, and 76%/53% of contacts *y* and *s* could potentially form Coulombic interactions. For the α-helical fold, six intrahelical hydrophobic contacts and one set each of interhelical contacts, strand-helix contacts, and helix-capping contacts were observed (**Figure 3C**). Overall, 96% of interhelical contacts were hydrophobic, 94% of helix-capping residues could potentially form an i-4→i or i→i backbone-to-sidechain hydrogen bond, 85% of residues in the helix-loop interaction had a charged residue in one position, but not both, and 80% of residues in intrahelical contact *a* were both hydrophobic. The remaining contacts gave more mixed results, perhaps due to hydrophobic residues contacting the hydrophobic portion of their hydrophilic partners. Previous work has shown that interdomain interactions also contribute significantly to RfaH fold switching (*16*). Unfortunately, these interactions could not be identified by coevolutionary analysis (**Figure S9**), a likely result of the limited number of JPred-filtered sequences available.

Our results suggest that the sequences of fold-switching CTDs poise them to assume two disparate folds. Thus, it might be reasonable to expect these sequences to be relatively homogeneous, especially since variants of another fold switcher, human XCL1, lose their ability to switch folds below a relatively high identity threshold (60%, (*28*)). The opposite is true. Sequences of putative fold-switching CTDs are significantly more heterogeneous (20.4% mean/19.4% median sequence identity) than sequences of predicted single folders (40.5% mean/42.5% median sequence identity, **Figure 4A**). Accordingly, among the sequences tested experimentally, similar mean/median sequence identities were observed: 21.0%/21.1% (fold switchers), 43.2%/41.2% (single folders, **Figure 4B**). Additionally, fold-switching CTDs were predicted in most bacterial phyla, and many were predicted in archaea and eukaryotes as well (**Figure 4C, Data S2**). These results suggest that many highly diverse CTD sequences can switch folds between an α-helical hairpin and a β-roll in organisms from all kingdoms of life.

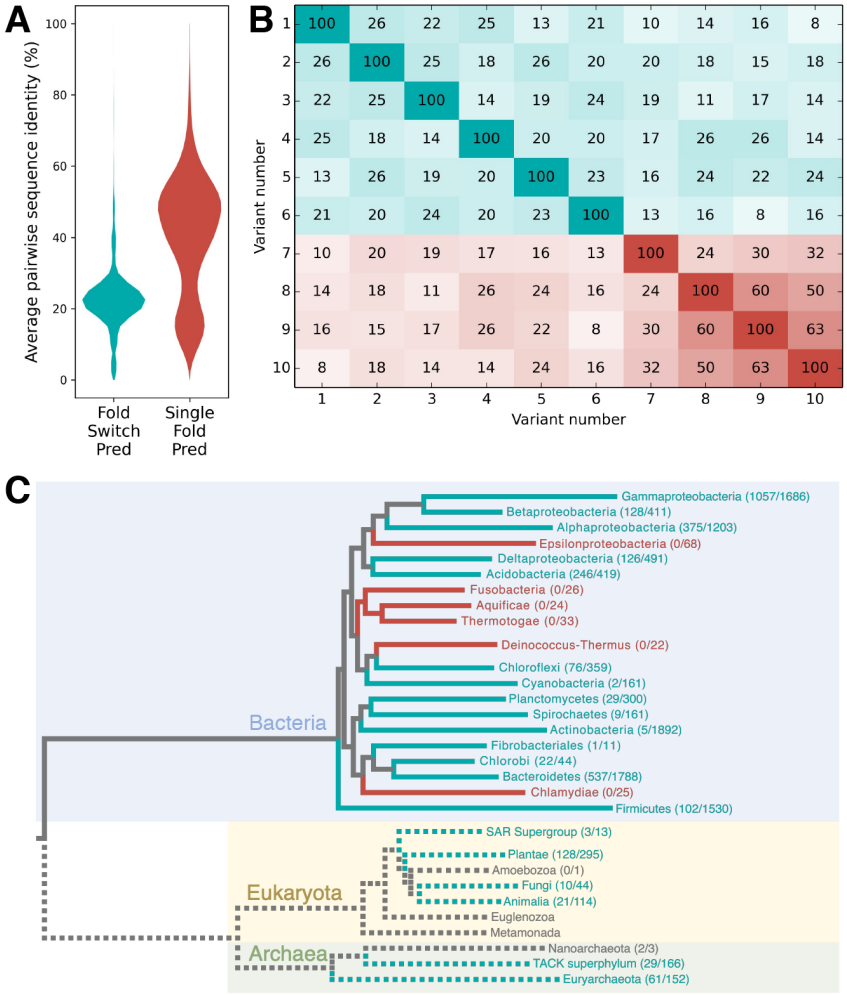


Fig. 4. The sequences of fold-switching CTDs are highly diverse and found in a wide variety of bacterial phyla. (A) Violin plots of pairwise sequence identities differ significantly for putative fold switchers and putative single folders. On average, pairwise sequence identities are lower for putative fold switchers (20.4%) than single folders (40.5%). **(B)** Sequence identity matrix of the CTDs of variants 1-10 in **Figure 2**. Numbers in each box represent the % identity of the two variants compared. Darker boxes represent higher identity levels. **(C)** Fold-switching CTDs are predicted in many bacterial phyla and other kingdoms of life. Numbers next to taxa represent #predicted fold switchers/#total sequences. Gray branches represent unidentified common ancestors, since the evolution of fold-switching NusGs is unknown. Dotted lines represent lower-confidence predictions since fold switching has not been confirmed experimentally in archaea and eukaryota. Fold-switching/single-folding predictions are represented by teal/red colorings; predictions in branches with fewer than 10 sequences are gray.

Why might the sequence diversity of fold-switching CTDs exceed those of single folders? Functional diversity is one likely explanation (*29*). Previous work has shown that NusG^{SP}s drive the expression of diverse molecules from antibiotics to toxins (*15*). Our method predicts that many of these switch folds. Furthermore, since helical contacts are conserved among at least some fold-switching CTDs, it may be possible that CTD sequence variation is less constrained in other function-specific positions. The fold-switching mechanism of RfaH allows it to both regulate transcription and expedite translation, presumably quickening the activation of downstream genes. NusG^{SP}s are likely under strong selective pressure to conserve this mechanism when the regulated products control life-or-death events, such as the appearance of rival microbes or desiccation. Supporting this possibility, RfaH, LoaP, and UpxY usually drive operons controlling rapid response to changing environmental conditions such as macrolide antibiotic production (*22*), antibiotic-resistance gene expression (*15*), virulence activation (*30*), and biofilm formation (*31*).

Our method was sensitive enough to predict fold-switching proteins, setting it apart from other state-of-the-art methods. These other methods assume that all homologous sequences adopt the same fold, as evidenced by their use of sequence alignments that contained both fold-switching

and single-folding sequences. These mixed sequence alignments biased their predictions. While those predictions are partially true since both fold-switching and single-folding CTDs can fold into β-rolls, they miss the alternative helical hairpin conformation and its regulatory function (*14*). Computational approaches that account for conformational variability and dynamics, a weakness in even the best predictors of protein structure (*9*), could lead to improved predictions. This need is especially acute in light of recent work showing how protein structure is influenced by the cellular environment (*32*), and it could inform better design of fold switchers, a field that has seen limited success (*33-35*).

Our results indicate that fold switching is a pervasive, evolutionarily conserved mechanism. Specifically, we predicted that 25% of the sequences within a ubiquitous protein family switch folds and found that residue-residue contacts unique to each fold-switching conformation are conserved through evolution. This sequence-diverse dual-fold conservation challenges the one-sequence-one-structure protein folding paradigm and indicates that foundational principles of protein structure prediction may need to be revisited.

The success of our method in the NusG superfamily suggests that it may have enough predictive power to identify fold switching in protein families believed to contain only single folders. Such predictions would be particularly useful since many fold switchers are associated with human disease (*3-6*). Given the unexpected abundance of fold switching in the NusG superfamily, there may be many more unrelated fold switchers to discover.

References and Notes:

1. L. L. Porter, L. L. Looger, Extant fold-switching proteins are widespread. *Proc Natl Acad Sci U S A* **115**, 5968-5973 (2018).
2. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).
3. A. K. Kim, L. L. Porter, Functional and Regulatory Roles of Fold-Switching Proteins. *Structure* **29**, 6-14 (2021).
4. B. P. Li *et al.*, CLIC1 Promotes the Progression of Gastric Cancer by Regulating the MAPK/AKT Pathways. *Cell Physiol Biochem* **46**, 907-924 (2018).
5. D. Giganti *et al.*, Secondary structure reshuffling modulates glycosyltransferase function at the membrane. *Nat Chem Biol* **11**, 16-18 (2015).
6. D. E. Gordon *et al.*, Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370**, (2020).
7. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072-1080 (2012).
8. J. Yang *et al.*, Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A* **117**, 1496-1503 (2020).
9. A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).
10. M. Lopez-Pelegrin *et al.*, Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase. *Angew Chem Int Ed Engl* **53**, 10624-10630 (2014).
11. N. Chen, M. Das, A. LiWang, L. P. Wang, Sequence-Based Prediction of Metamorphic Behavior in Proteins. *Biophys J* **119**, 1380-1390 (2020).

12. A. K. Kim, L. L. Looger, L. L. Porter, A high-throughput predictive method for sequence-similar fold switchers. *Biopolymers*, e23416 (2021).
13. S. Mishra, L. L. Looger, L. L. Porter, Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci* **28**, 1487-1493 (2019).
14. J. Y. Kang *et al.*, Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650-1662 e1614 (2018).
15. B. Wang, V. M. Gumerov, E. P. Andrianova, I. B. Zhulin, I. Artsimovitch, Origins and Molecular Evolution of the NusG Paralog RfaH. *mBio* **11**, (2020).
16. B. M. Burmann *et al.*, An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **150**, 291-303 (2012).
17. P. K. Zuber, K. Schweimer, P. Rosch, I. Artsimovitch, S. H. Knauer, Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun* **10**, 702 (2019).
18. A. Drozdetskiy, C. Cole, J. Procter, G. J. Barton, JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* **43**, W389-394 (2015).
19. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
20. G. A. Belogurov *et al.*, Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell* **26**, 117-129 (2007).
21. C. Wang *et al.*, Structural basis of transcription-translation coupling. *Science* **369**, 1359-1365 (2020).
22. J. R. Goodson, S. Klupt, C. Zhang, P. Straight, W. C. Winkler, LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. *Nat Microbiol* **2**, 17003 (2017).
23. B. Wang, I. Artsimovitch, NusG, an Ancient Yet Rapidly Evolving Transcription Factor. *Front Microbiol* **11**, 619618 (2020).
24. D. E. Kim, D. Chivian, D. Baker, Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526-531 (2004).
25. T. A. Hopf *et al.*, The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582-1584 (2019).
26. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015).
27. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
28. A. F. Dishman *et al.*, Evolution of fold switching in a metamorphic protein. *Science* **371**, 86-90 (2021).
29. N. Tokuriki, D. S. Tawfik, Protein dynamism and evolvability. *Science* **324**, 203-207 (2009).
30. J. A. Leeds, R. A. Welch, RfaH enhances elongation of Escherichia coli hlyCABD mRNA. *J Bacteriol* **178**, 1850-1857 (1996).
31. C. Beloin *et al.*, The transcriptional antiterminator RfaH represses biofilm formation in Escherichia coli. *J Bacteriol* **188**, 1316-1331 (2006).
32. W. B. Monteith, R. D. Cohen, A. E. Smith, E. Guzman-Cisneros, G. J. Pielak, Quinary structure modulates protein stability in cells. *Proc Natl Acad Sci U S A* **112**, 1739-1742 (2015).

33. P. A. Alexander, Y. He, Y. Chen, J. Orban, P. N. Bryan, A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A* **106**, 21149-21154 (2009).
34. X. I. Ambroggio, B. Kuhlman, Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **128**, 1154-1161 (2006).
35. K. Y. Wei *et al.*, Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proc Natl Acad Sci U S A* **117**, 7208-7215 (2020).
36. C. UniProt, The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142-148 (2010).
37. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
38. F. Pedregosa, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. Perrot, M. and Duchesnay, E., Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).
39. P. J. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
40. N. P. Brown, C. Leroy, C. Sander, MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380-381 (1998).
41. A. A. Hagberg, Schult, D. A., and Swart, P. J., in *Proceedings of the 7th Python in Science Conference,* T. V. Gäel Varoquaux, Jarrod Millman, Ed. (Pasadena, CA USA, 2008), pp. 11-15.
42. B. Rost, Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94 (1999).
43. B. Ruan, K. E. Fisher, P. A. Alexander, V. Doroshko, P. N. Bryan, Engineering subtilisin into a fluoride-triggered processing protease useful for one-step protein purification. *Biochemistry* **43**, 14539-14546 (2004).
44. A. Micsonai *et al.*, BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res* **46**, W315-W322 (2018).
45. S. B. Azatian, N. Kaur, M. P. Latham, Increasing the buffering capacity of minimal media leads to higher protein yield. *J Biomol NMR* **73**, 11-17 (2019).
46. M. Cai, Y. Huang, R. Yang, R. Craigie, G. M. Clore, A simple and robust protocol for high-yield expression of perdeuterated proteins in Escherichia coli grown in shaker flasks. *J Biomol NMR* **66**, 85-91 (2016).
47. J. Marley, M. Lu, C. Bracken, A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR* **20**, 71-75 (2001).
48. F. Delaglio *et al.*, NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-293 (1995).
49. J. Ying, F. Delaglio, D. A. Torchia, A. Bax, Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J Biomol NMR* **68**, 101-118 (2017).
50. W. Lee, M. Tonelli, J. L. Markley, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325-1327 (2015).

51. Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* **44**, 213-223 (2009).
52. T. Kortemme, A. V. Morozov, D. Baker, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**, 1239-1259 (2003).
53. R. Srinivasan, G. D. Rose, A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* **96**, 14258-14263 (1999).
54. M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175 (2011).
55. C. R. Harris *et al.*, Array programming with NumPy. *Nature* **585**, 357-362 (2020).
56. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*, (2021).
57. A. Rambaut. (2012).
58. W. Shen, H. Ren, TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics*, (2021).
59. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
60. J. D. Hunter, Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**, 90-95 (2007).
61. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

Acknowledgements:

L.L.P. thanks Drs. Marius Clore and Carolyn Ott for constructive input throughout this project and for many excellent suggestions concerning the text and figures. Thanks to Dr. Nico Tjandra for generously sharing his lab and office space and Dr. David Landsman for administrative support. We also thank Drs. Liskin Swint-Kruse, Gisela Storz, Nico Tjandra, David Nyenhuis, and Daniel Morris for helpful comments concerning the text, Dr. Christos Kougentakis for helping to collect NUS NMR data, Dr. Nicholas Fitzkee for sharing his isotope labeling protocol, Dr. Aaron Robinson for help with NMR samples, and Dr. Xiaofang Jiang for recommending TaxonKit and ITOL. This work utilized resources from the NHLBI Biophysics Core, the NHLBI Protein Expression Facility, and the NIH HPS Biowulf cluster (http://hpc.nih.gov), and it was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health and Howard Hughes Medical Institute.

Materials and Methods

Identification of NusG-like sequences

NusG-like sequences were identified from the October 2019 Uniprot90 (*36*) database using an iterative BLAST (*19*) approach. Specifically, the *E. coli* RfaH sequence (Uniprot ID Q0TAL4) was BLASTed against the database. All hits with a maximum e-value of 10^{-4} were aligned using Clustal Omega (*37*), which generated their sequence identity matrices from the resulting alignment. Sequences were clustered by their identities using the agglomerative clustering algorithm from the python module scikit-learn (*38*). Sequence identity between proteins in each cluster was $\geq$ 78%. Randomly selected sequences from the 25 largest clusters were then individually BLASTed against the Uniprot90 database, and the resulting hits were combined; redundant identical hits from independent searches were removed. This procedure (search-align-

cluster) was repeated two additional times to generate the full list of 15,516 sequences in 305 clusters.

Determination of CTDs

Sequences of annotated RfaHs were aligned to the sequence of *E. coli* RfaH (Uniprot ID Q0TAL4) using Clustal Omega (*37*). CTDs were defined as up to 50 residues, but not shorter than 40 if the CTD region comprised <50 residues, beginning with the positions that aligned to the RfaH sequence *KVIIT*. Sequences of proteins not annotated as RfaH were aligned to the *E. coli* NusG sequence (Uniprot ID P0AFG0) using Clustal Omega. CTDs were defined as 50 residues beginning with positions that aligned the NusG sequence *EMVRV*. Because of their diversity, sequences from each individual cluster were aligned against the NusG sequence separately, each using Clustal Omega. The number of sequences with CTDs long enough to make these predictions totaled 15,195 (**Data S1**), 98% of all NusG-like sequences identified.

JPred4 predictions

JPred4 (*18*) predictions were carried out as in (*12*), sections 2.4 and 2.6. In further detail, they were first performed on all 50-residue CTD sequences using two databases: the JPred database (**http://www.compbio.dundee.ac.uk/jpred/about_RETR_JNetv231_details.shtml**) from 2014 and the Uniprot90 database from January 2021. Sequences of each prediction were aligned against the *E. coli* NusG sequence (beginning with *EMVRV*) using Biopython (*39*) Bio.pairwise2.localxs with gap opening/extension scores of -1.0/-0.5. Secondary structure predictions of the sequence in question and of *E. coli* NusG were reregistered according to the resulting pairwise alignments and compared as in (*12*). Predictions were considered high-confidence if at least 5 sequences were in the MView (*40*)-generated alignments used by JPred.

We found that the first 10 residues in these 50-residue sequences were similar enough to NusG CTDs that NusG-like sequences overwhelmed sequence alignments informing the predictions, and many likely fold-switching sequences were predicted to be single folders. To circumvent this problem, predictions from both databases were rerun on 40-residue sequences (starting with the first residue that aligned to *ADFNG*… for NusG sequences and *FQAIF*… for RfaH sequences). Predictions were made as with 50-residue sequences. All predictions reported in the main text were from 40-residue sequences, except those in **Figure 1B**.

Force-directed graph

The 305 clusters generated from all full-length NusG sequences were plotted on a force-directed graph using the *spring_layout* function from python NetworkX (*41*) with a spring constant of 0.3 and 1000 iterations. Nodes with ≥50% of sequences predicted to switch folds were colored teal; nodes with <50% of sequences predicted to switch folds were colored red. Nodes with no predictions were colored gray. Nodes 1 and 7 were colored differently from their average predictions (single folding, Node 1; fold-switching, Node 7) to highlight the prediction of the sequence validated experimentally, which differed from the average. Edges represented average pairwise identities between nodes ≥24%, a threshold taken from (*42*) for sequences of 162 residues (the length of *E. coli* RfaH).

Genomic analysis of sequences

The annotated genomes (protein .fasta and .gtf annotation) of 31,554 bacterial species were downloaded from Ensembl Bacteria in April 2021. Genomic annotation of NusG was defined as being within 10 kb of a gene annotated as either "SecE," "RplK," "RplA," or "ribosomal protein L11" by text matching. Most bacterial genomes are incompletely assembled and annotated – the genes were required to be within the same chromosome, contig, or plasmid. Each Uniprot sequence in the database of 15,516 was mapped to an Ensembl locus if the species was consistent, and if sequence identity was greater than 90%. Annotation was fetched from Ensembl, as well – this was usually, but not always, consistent with the Uniprot annotation.

Of the 15,516 Uniprot sequences, 7975 mapped to Ensembl genomes. Cursory analysis of some non-mapping sequences suggested that: 1) some Ensembl genomes had incomplete collation of all ORFs, and 2) there were frame-shifts and other errors in some Uniprot sequences and some Ensembl genomes. This was also the case for some of the sequences predicted to potentially be fold-switching NusGs: for instance, Uniprot entry A0A0T8ANM4 is frame-shifted relative to the Ensembl genome, producing a C-terminal sequence predicted to switch folds.

Of the 5,435 sequences that mapped to Ensembl loci with *SecE*/*RplK*/*RplA* within 10kb, only 22 had a separation of >1kb, and only 59 had a separation of >270bp – this set of 59 includes 4 proteins predicted to be fold-switching, one of which is a verified RfaH from (*15*), indicating that a shorter threshold of distance to *SecE*/*RplK*/*RplA*, perhaps coupled with determining distance several other conserved *NusG*-*SecE* operon genes, could reduce the false-positive rate caused by mistakenly annotating $NusG^{SP}$s as housekeeping NusGs.

For a small number of sequences that mapped to qualitatively dissimilar genes (e.g., one genomically consistent as being a NusG, another not), the 2nd mapping is given in **Data S1**, beginning in column AH.

Additionally, of the 600 RfaH sequences that mapped to an annotated Ensembl locus, only one fell within a NusG-like operon (~7kb away).

<u>Expression and purification of variants 1-16</u>

All variants were ordered from IDT as gBlocks. Except for variant #8, these variants were digested with HindIII and EcoRI and ligated into the pPAL7 vector (Bio-Rad) with an N-terminal 6-His tag cloned using a Q5 mutagenesis kit (New England Biolabs). Variants were transformed into *E. coli* BL21-DE3 cells (New England Biolabs), grown in LB at 37° to an OD_{600} of 0.6-0.8, after which they were incubated at 20°C for 30 minutes, induced with 0.1 mM IPTG, and grown overnight, shaking at 225-250 rpm. Variant #8 was cloned into the same vector as the other variants using In-Fusion and expressed as the other variants but at 18°C instead of 20°C. The cells from all cultures were pelleted at 10,000*xg* for 10 minutes at 4°C, resuspended in 2 mL lysis buffer (50 mM Tris, 150 mM NaCl, 5% glycerol, 1 mM DTT, 10 mM imidazole, pH 8.7) and frozen at -80°C for later purification. Sequencing of all variants was verified by Psomagen.

Thawed cell pellets were resuspended in 25 mL lysis buffer per 1 L of culture grown. 100 mg of DNAseI, 5 mM $CaCl_2$, 5 mM $MgSO_4$ and 1/2 of a cOmplete EDTA-free protease cocktail inhibitor tablet (Roche) were added per 25 mL of lysis buffer. Cells were lysed by 2 passes through an EmulsiFlex-C3 homogenizer (Avestin). The homogenized lysate was centrifuged for 45 minutes at 40,000*xg* at 4°C, and its soluble fraction was loaded immediately onto either a 1 mL Ni column (GE HisTrap HP) or an Econo-Pac (Bio-Rad) gravity column with 0.5-1 mL IMAC Ni Resin (Bio-Rad). Soluble lysate was loaded on ice for the HisTrap column, and gravity columns were loaded

and kept at 4°C. The HPLC Ni columns were washed with 100 mM phosphate and 500 mM NaCl, pH 7.4, equilibrated in 100 mM phosphate, pH 7.4, and eluted by gradient with 0.5 M imidazole, 100 mM phosphate, pH 8.0 at 2 mL/minute on an ÄKTA Avant. The gravity columns were washed and equilibrated with 10 column volumes each of the same buffers, and protein was eluted at 3 different imidazole concentrations: 100 mM, 500 mM and 2M, all in 100 mM phosphate, pH 7.4.

Nickel-purified samples were then loaded onto 1- or 5-mL Profinity eXact (*43*) columns (Bio-Rad), washed twice with one column-volume of 2M NaOAc, and eluted with 100 mM phosphate, 10 mM azide, pH 7.4 at 0.2 mL/minute. Cleavage kinetics for some variants (1, 4, and 6) were too slow to get adequate tagless protein. In these cases, columns were equilibrated with 100 mM phosphate, 10 mM azide, pH 7.4 overnight at 4°C. Tagless protein was concentrated in 10 kDa MWCO concentrators (Millipore), and the buffer was exchanged to 100 mM phosphate, pH 7.4. A small amount of high-molecular-weight impurity (<10% of the sample) from variants #1 and #4 was removed by running the tagless sample through a 50 kDa MWCO concentrator (Millipore) and keeping the low molecular weight fraction that passed through the filter. Sample purities were assessed by gel electrophoresis (Thermofisher NuPAGE 4-12% Bis-Tris gels, Thermofisher MES buffer, Bulldog Bio Coomassie Stain), and concentrations were measured on a NanoDrop OneC (Thermo Scientific).

Circular dichroism (CD) spectroscopy

All CD spectra were collected on Chirascan spectrometers (Applied Photophysics) in 1 mm quartz cuvettes in 100 mM phosphate, pH 7.4. Protein concentrations ranged from 8-12 mM, and scan numbers ranged from 5-10. Scans were averaged, and averaged baselines of buffer-blank 1 mm cuvettes were subtracted from the spectra. The resulting spectra were entered into the BestSel (*44*) webserver (https://bestsel.elte.hu/index.php), so that their ratio of helix (helix+distorted helix):strand (parallel+antiparallel) could be computed.

CTDs of variants #5 and #8

Full-length variants #5 and #8 were shortened to 64 and 68 residues, respectively, using Q5 mutagenesis (New England Biolabs). Their sequences were confirmed by Sanger sequencing (Psomagen) and are reported in **Table S2.**

Expression and purification of NMR samples

Based on the protocols in (*45-47*), BL-21 DE3 cells (New England Biolabs) expressing all NMR samples were grown in LB to an OD_{600} of 0.6 and pelleted at 5000*xg* for 30 minutes at 4°C. The pellets were resuspended in 1X M9 at half of the initial culture volume and pelleted at 5000*xg* for 30 minutes at 4°C. Pellets were then resuspended at ¼ initial culture volume in 2X M9, pH 7.0-7.1, 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, with 1 g $^{15}NH_4Cl$/L, and 4 g of either unlabeled or ^{13}C-labeled glucose (Cambridge Isotope Laboratory)/L and equilibrated at 20°C for 30 min, shaking at 225 rpm, then induced with 1 mM IPTG and grown overnight. Cells were pelleted at 10,000xg for 10 minutes at 4°C. All labeled variants were purified by FPLC (ÄKTA Avant 25) using the same methods as variants 1-16 above in 5 mL HisTrap HP columns (Cytiva) and 5 mL Profinity eXact columns (BioRad).

^{1}H-^{15}N HSQCs of variants #5 and #8

All spectra were collected on Bruker 600 MHz spectrometers equipped with z-gradient cryoprobes and processed with NMRPipe (*48*). Variant #8 (full-length and CTD) and variant #5 CTD HSQCs were collected in 100 mM phosphate, pH 7.4 at 298 K. Under those conditions the spectrum of full-length variant 5 was broad, even with 1 mM DTT added, but peaks narrowed upon changing the buffer conditions to 25 mM HEPES, 50 mM NaCl, 5% glycerol, 1 mM DTT, pH 7.5, and collecting the spectrum at 303K. Protein concentrations ranged from 100-300 uM.

Assignments of KCTD and TCTD

^{13}C-labeled 5CTD and 8CTD were expressed and purified as above. For each variant, HNCACB, CBCA(CO)NH, and HNCO experiments were collected on Bruker 600 MHz spectrometers with cryoprobes. Spectra of 8CTD (80 uM) were collected using nonuniform sampling and were processed with SMILE (*49*). All NMR spectra were processed using NMRpipe (*48*). Resonances were assigned manually with NMRfam Sparky (*50*), and secondary structures were determined using TALOS+ (*51*).

Coevolutionary analysis

Structure predictions of the 6 fold-switching variants were calculated by entering their full-length sequences (**Table S3**) into the EVCouplings (*25*), Robetta (*24*), and Phyre2 (*26*) webservers (https://evcouplings.org, https://robetta.bakerlab.org, http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). EVCouplings predictions with the recommended e-value cutoffs for chosen: (Variant 1: e-3, 2: e-5, 3: e-5, 4: e-20, 5: E-5, 5: e-5). High-confidence predictions for shorter sequences of 40 or 50 residues could not be obtained from either EVCouplings or Robetta. Predicted residue-residue contacts of *E. coli* RfaH from EVCouplings/Robetta with probabilities ≥ 99%/92% were plotted in **Figure 3A**, and residue-residue contacts from GREMLIN (*27*) with probabilities ≥90% were plotted in **Figure 3B**. These thresholds were determined by maximizing the ratio of true positives to false positives. True positives were considered to be couplings with heavy atoms within 5.0 Å in either the 2OUG or the 2LCL crystal structures where at least one of the 2 heavy atoms was from a side chain; one additional contact between residues 140 and 151 was added because they were separated by 5.2 Å within the NMR structure and therefore likely within error of 5.0Å. Contacts were considered hydrophobic if both atoms in contact were hydrophobic, Coulombic if two atoms in contact had opposite charge and C-N-O/C-O-N angles ≥ 90°, and helix caps if the distance between sidechain donor/acceptor ≤4° and C-N-O/C-O-H angles ≥ 90° (*52*). All distances and angles were calculated using LINUS (*53*).

CTD sequences for GREMLIN webserver (http://gremlin.bakerlab.org/submit.php) analysis in **Figure 3B** were obtained by clustering all JPred predictions by Affinity Propagation using the python Scikit-learn module (*38*) with damping of 0.99 and a maximum number of 10,000 iterations. Affinities were precomputed by comparing each 40-residue prediction position-by-position, with the following scores: identical predictions (EE,HH,--): 0, coil:secondary structure discrepancies (H-,E-,-H,-E): 0.5, and helix:strand discrepancies (HE,EH): 10, and selecting the cluster with the sequence of *E. coli* RfaH (639 sequences). These sequences were aligned with Clustal Omega and inputted into GREMLIN. 4 iterations of HHBlits (*54*) were run on the initial alignment with E-values of 10^{-10}. Coverage and remove gaps filters were both set to 75.

GREMLIN webserver analyses were run on EVCouplings and Robetta multiple sequence alignments seeded with the sequence of *E. coli* RfaH. These alignments were taken from EVCouplings *align* and Robetta .msa.npz files. No additional iterations of HHPred were run on either alignment. Coverage and remove gaps filters were both set to 75.

Pairwise sequence identities

Pairwise sequence identity matrices of predicted fold-switching/single-folding CTDs were calculated using Geneious. The alignments for these sequences were first manually curated to remove sequences that did not align well with the majority; manually curated alignments retained at least 98% of all sequences. The mean/median sequence identities of these two groups were determined from the upper triangular matrices of each matrix, excluding positions of identity, using numpy (*55*). Pairwise sequence identity matrices of the CTDs of the 10 variants were determined with Clustal Omega.

Phylogenetic tree

The tree in **Figure 4C** was generated by downloading the Interactive Tree of Life (*56*) (https://itol.embl.de/itol.cgi), loading it into FigTree (*57*), and collapsing branches at the phyletic level, except for Proteobacteria, which were left at the class level because of recent phylogenetic work on proteobacterial RfaH (*15*).

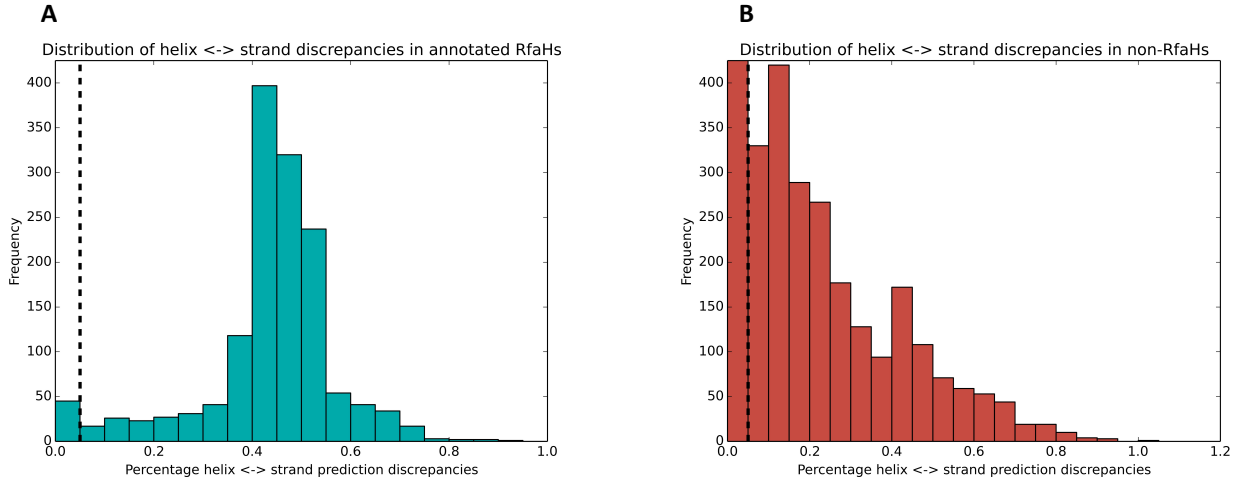
Bacterial species from each NusG sequence were obtained from their Uniprot headers. These species were mapped to their respective phyla using TaxonKit (*58*) and matched with their predictions. Phyla with fold-switching/single-folding predictions were listed using a python script, and branches of the tree were then colored manually in Adobe Illustrator.

Eukaryotic and archaeal NusG homologs were obtained by running 3 rounds of PSI-BLAST on the nr database with the following seed sequences: L1IE32, A0A0N95N5M7, UPI0005F5777A, A0A2E6HKN0. Redundant sequences were removed using CD-HIT (*59*) at a 98% sequence identity threshold (at least 1 amino acid difference).

Figures

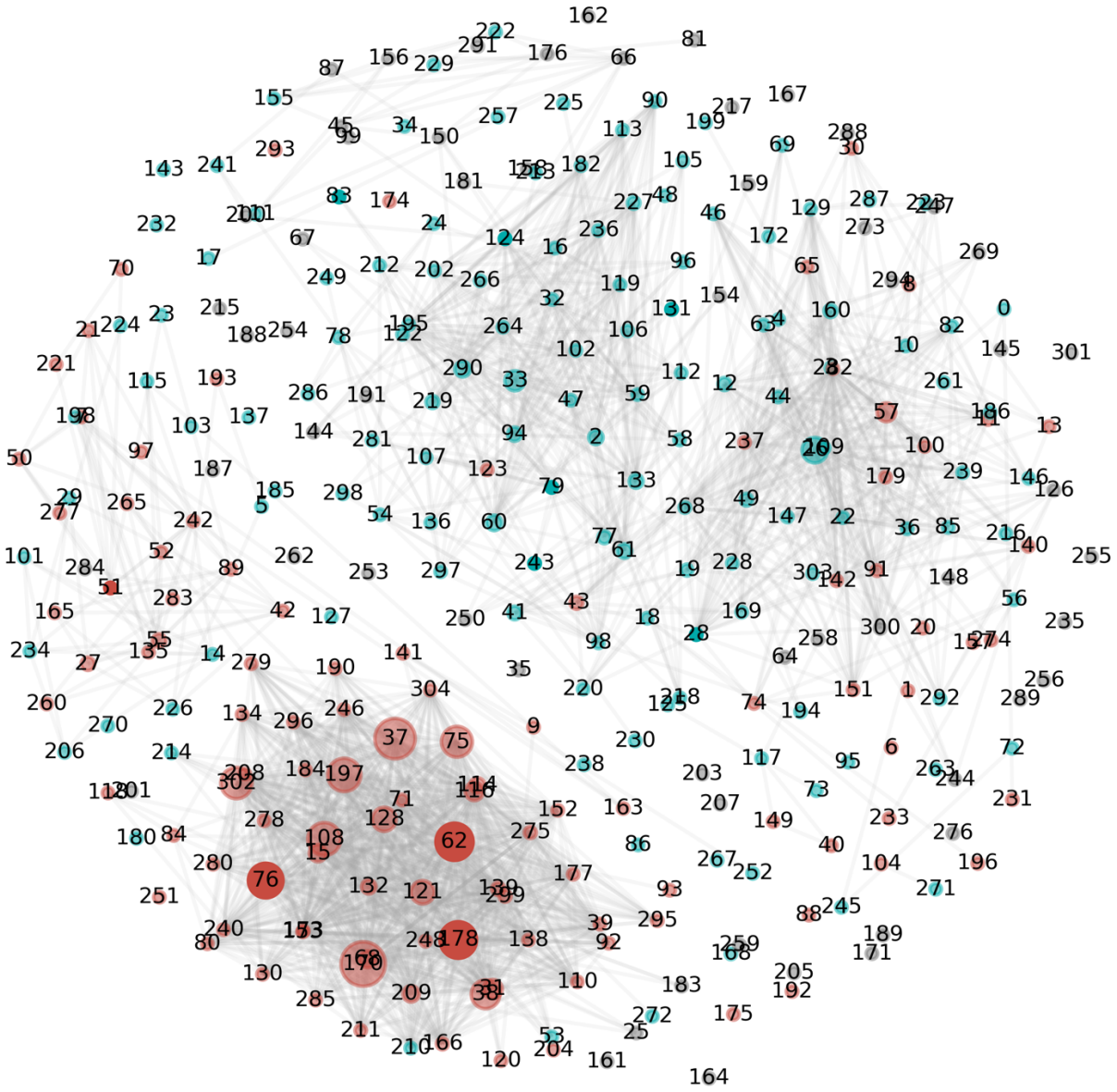
Figures 1C, 2A, 2B, 3A, 3B, 4A, and 4B were generated using Matplotlib (*60*). The figures of all protein structures (Figures 1A and 3C) were generated using PyMOL (*61*).

Fig. S1.



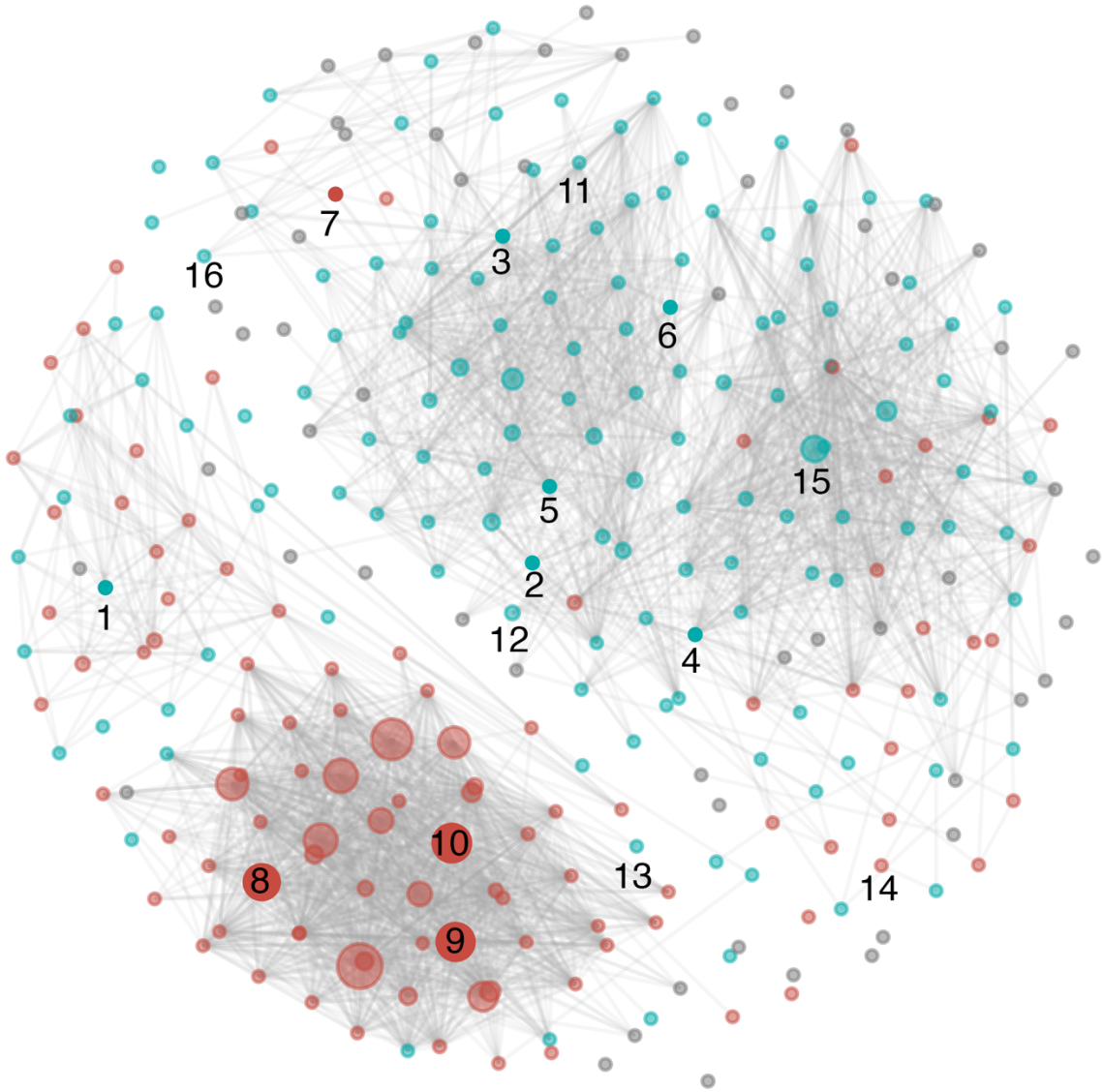
Distributions of prediction discrepancies for Uniprot-annotated RfaHs (A) and sequences not annotated as RfaH (B). Black dotted line is the cutoff point for fold-switching and single-folding predictions: predictions with ≥5% discrepancy to the *E. coli* NusG sequence were predicted to switch folds. This cutoff was taken from Kim, et al. (*12*). Because of the low threshold, experiments were performed on constructs just above/below the threshold (Constructs 1, 4, and 7, respectively, **Table S1, Figure 2**). For comparison, the y-axis of both plots was limited to 410 counts. All bins fell below that threshold except for the first bin of panel B, which had 11,032 counts (and thus is not shown to scale).

Fig. S2.



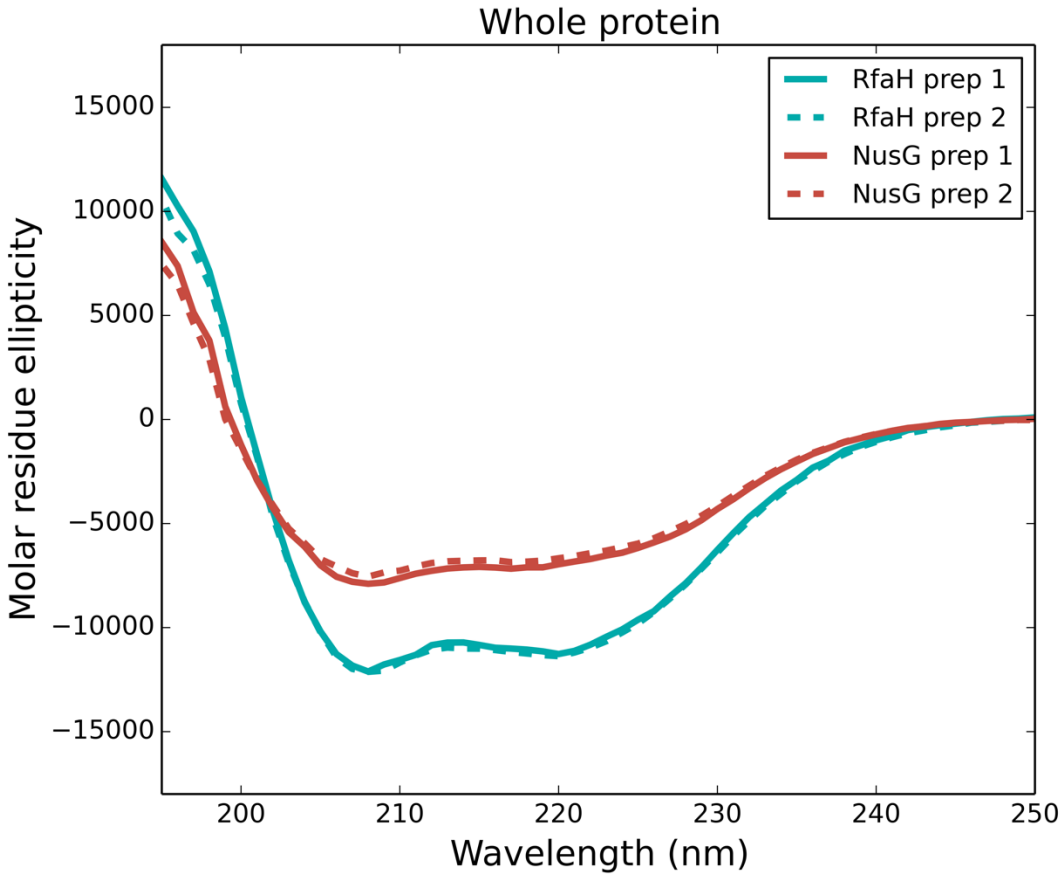
Sequence space diagram with cluster numbers labeled. Numbers correspond to the Cluster IDs (column 3) in **Data S1.**

Fig. S3.



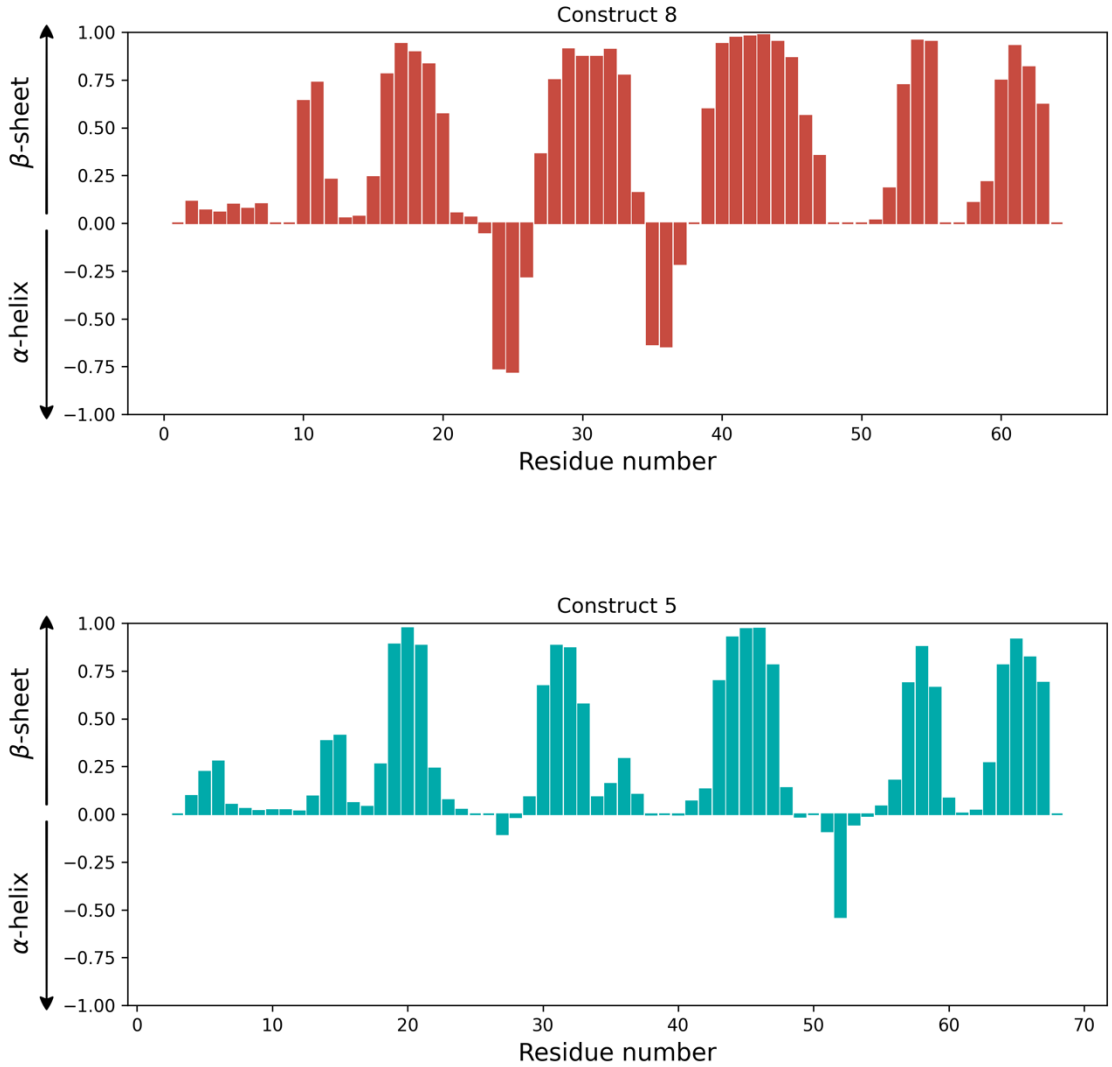
All sequence space constructs tested. Constructs 1-10 are labeled as in Figure 2. Constructs 11-16 were tested, but CD spectra could not be obtained because they did not express (Constructs 11-13 and 15-16) or because they were insoluble (Construct 14). With the exceptions of Constructs 8-10, all labels are directly below the nodes from which sequences were selected. Teal/red nodes: predicted to/not to switch folds on average; no high-confidence predictions were made for gray nodes. More information about each construct can be found in **Table S1**. Nodes 1 and 7 were colored differently from their average predictions (single folding, Node 1; fold-switching, Node 7) to highlight the prediction of the sequence validated experimentally, which differed from the average.

Fig. S4.



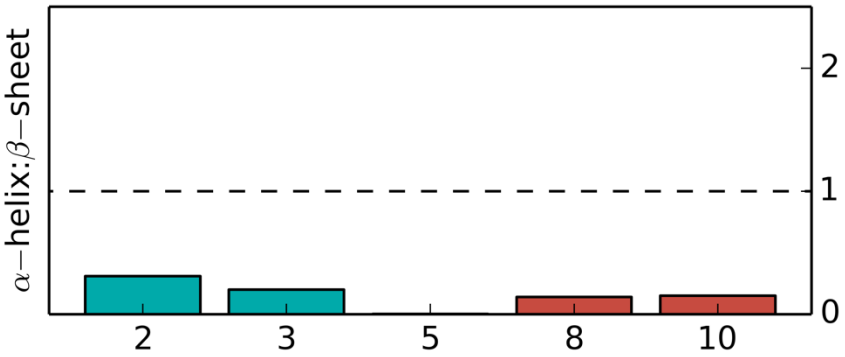
Circular dichroism (CD) spectra of two different *E. coli* RfaH preps differ significantly from two different *E. coli* NusG preps. By contrast, CD spectra of the two different preps of both RfaH and NusG are nearly identical to one another.

Fig. S5.



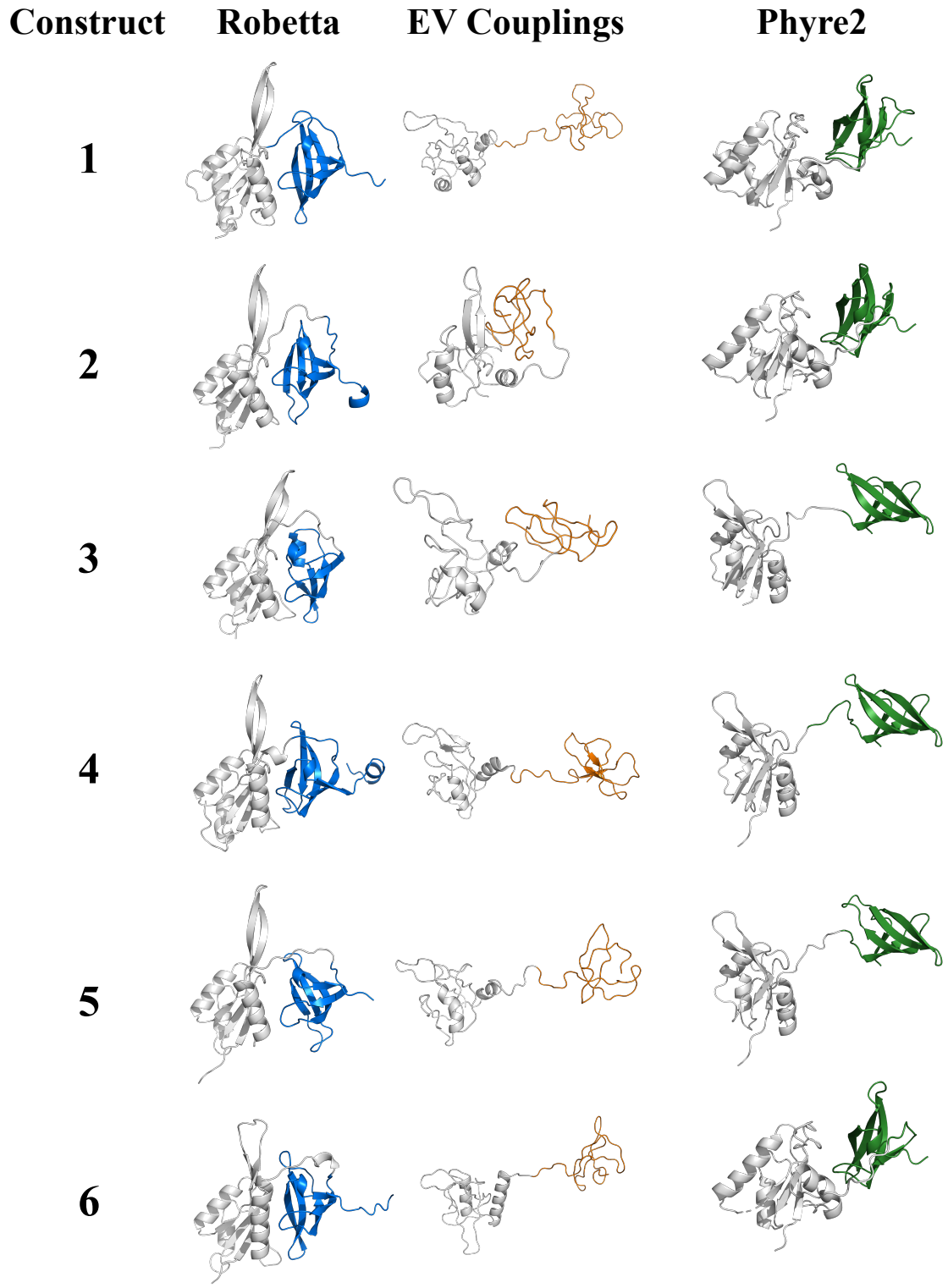
TALOS+ secondary structure predictions of the assigned CTDs of Construct 8 (above) and Construct 5 (below). Plots suggest that both CTDs fold into structures with 5 β-sheets, consistent with the NusG β-roll fold.

Fig. S6.



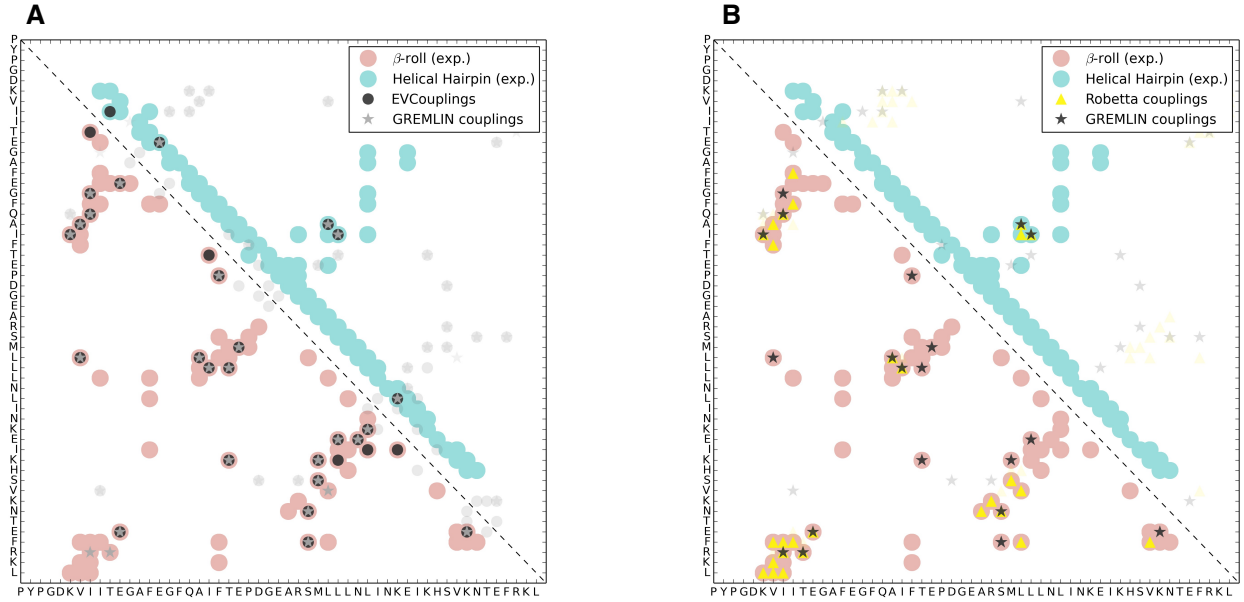
CD spectra of 5 CTDs fold predominantly into β-sheets. All variants were estimated to have 27.3% (2)-36.2% (3) β-strand content, while α-helical content ranged from 0.03% (5)-8.4% (2). Like variants 2 and 3, variant 5 appears to switch folds. Secondary structure content was estimated by the BestSel server (*44*). Variant numbers correspond to those in **Figure 2**.

Fig. S7.



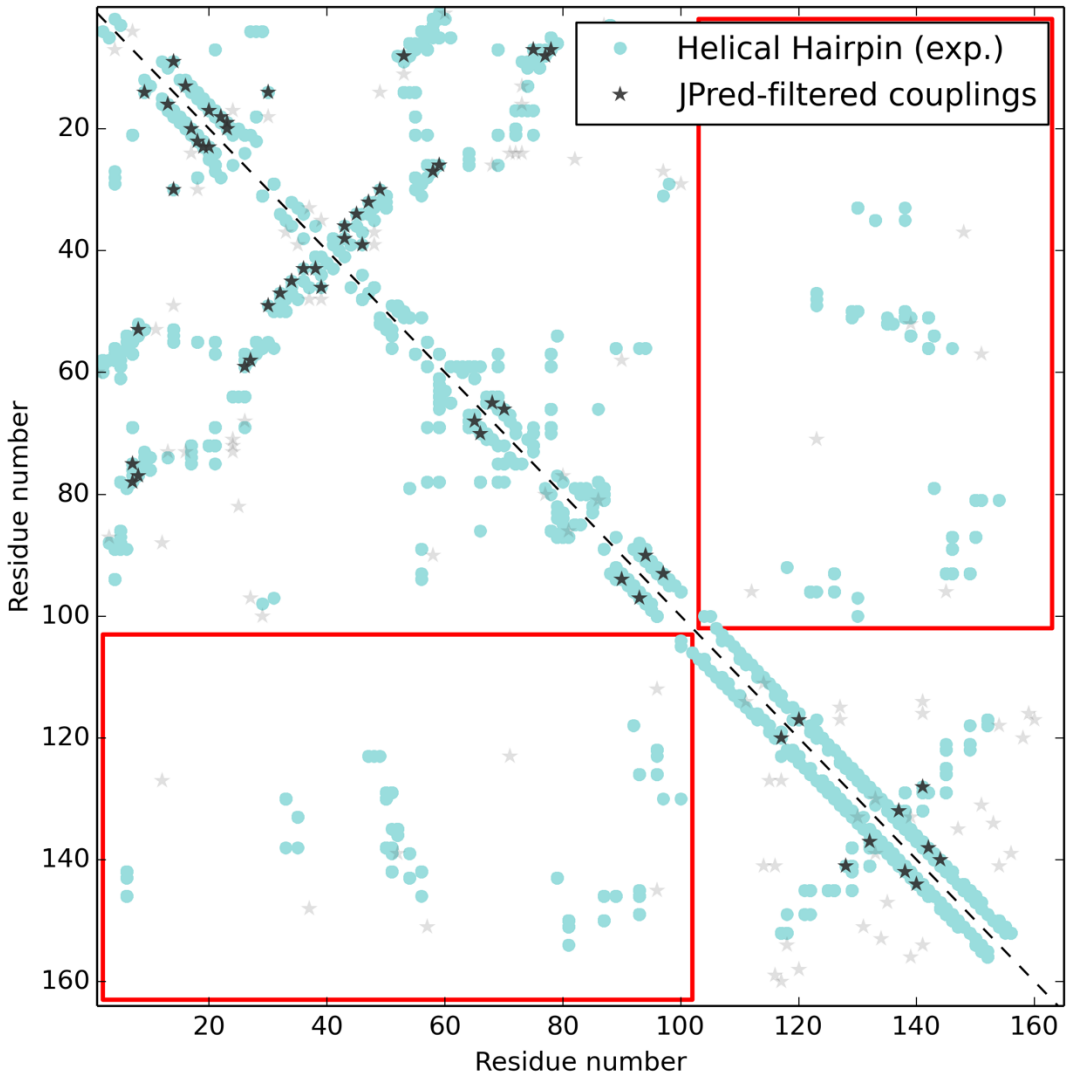
CTDs of the lowest-energy models for 6 proteins with RfaH-like folds (helical hairpin) are predicted assume β-sheet folds, including *E. coli* RfaH (Construct 3), which has an experimentally validated structure. CTDs are colored blue (Robetta), orange (EVCouplings), green (Phyre2)

Fig. S8.



GREMLIN couplings calculated from EVCouplings (A) and Robetta (B) sequence alignments largely match contacts from the experimentally determined β-roll fold (red, PDB ID 2LCL) but did not match any contacts unique to the helical hairpin fold (teal, PDB ID 2OUG_A).

Fig. S9.



JPred-filtered couplings of putative full-length fold switchers calculated by GREMLIN. No interdomain contacts (found within red boxes) were consistent with the experimentally determined structure of full-length RfaH (PDB ID: 2OUG).

Table S1.

#	ID	Cluster	Annotation	Phylum/Class	Pred	%H<->E	Express?	Soluble?
1	UPI000E4E22B5	51	LoaP	Firmicutes	FS	9%	Y	Y
2	A0A0S4NBF0	243	RfaH	Candidatus Kryptonia	FS	53%	Y	Y
3	Q0TAL4	124	RfaH	Gammaproteobacteria	FS	44%	Y	Y
4	B3EDK9	28	NusG	Chlorobi	FS	7%	Y	Y
5	A0A2J6WKD6	79	NGN domain-containing protein	Deferribacteres	FS	81%	Y	Y
6	E8N6B2	131	Putative RfaH	Chloroflexi	FS	50%	Y	Y
7	Q9F769	83	UpbY	Bacteroidetes	NFS	2%	Y	Y
8	A0A0P1LTF1	76	NusG	Candidatus Kryptonia	NFS	0%	Y	Y
9	P0AFG0	178	NusG	Gammaproteobacteria	NFS	0%	Y	Y
10	A0A1W1XWG8	62	NusG	Deltaproteobacteria	NFS	0%	Y	Y
11	*A1VS04*	*105*	*NusG*	*Betaproteobacteria*	*FS*	*65%*	*N*	-
12	*A0A1W1XVU0*	*41*	*NusG*	*Deltaproteobacteria*	*NFS*	*0%*	*N*	-
13	*A0A348AQW0*	*86*	*RfaH*	*Firmicutes*	*FS*	*30%*	*N*	-
14	*Q984H9*	*40*	*Mlr7998*	*Alphaproteobacteria*	*FS*	*38%*	*Y*	*N*
15	*A0A1M6KQH4*	*26*	*NusG*	*Bacteroidetes*	*FS*	*37%*	*N*	-
16	*A0A0F6QDM6*	*17*	*RfaH (on actX gene)*	*Gammaproteobacteria*	*FS*	*57%*	*N*	-

Table S2.

CTD Construct	Sequence
8	AELERIDVPF RVGDSVKVID GPFTDFSGVV QEVNSEKMKL KVMINIFGRK TPVELDFTQV EIEK
5	MVDGFIDTKS EEFKKGDTIL IKDGPFKDFV GIFQEELDSK GRVSILLKTL ALQPRITVDK DMIEKLHN

Table S3

Variant	Sequence
1	MMKPWYVLYVMGGKEQKILSLLNKGEDIKAFTPWKEVMHRVQGKRILVKKPLFPSYVFLE TELDPAVFHQKLMLYKSQINGILKELKYEDDISALHTEERAYLEGLMDEEHNVRLSKGEI LDGEVIITEGPLKGYESNIIRIDRHKRRAILNVRMNNQDLQVDVSLEIVKKIESQK
2	MDLNWYVLQTKPKQENLVESYLNLANIEVFNPKIQEIRYIGEKRKKITVLLFPCYVFAKL NPSLFDLVIYTRGVRKILGVNGRPKPIKESIIETIKERIRENSYIYVPENYEEFQLCQGD YVVVVDGPLKGFAGIVERINGSKAIVMLISMDYQVKADIPKFLLRKVDPEILE
3	MQSWYLLYCKRGQLQRAQEHLERQAVNCLAPMITLEKIVRGKRTAVSEPLFPNYLFVEFD PEVIHTTTINATRGVSHFVRFGASPAIVPSAVIHQLSVYKPKDIVDPSTPYPGDKVIITE GAFEGFQAIFTEPDGEARSMLLLNLINKEIKHSVKNTEFRKL
4	MKVTDRNSCWYAVYVRSRYEKKVHRMFLEKEVEAFLPLLETWRQWSDRKKKVSEPLFRGY VFVNIDMKAEHIKVLDTDGVVKFIGIGKTPSVISSRDIDWIKKLVREPDAVRRIVASLPP GQKVMVTAGPFKGLEGVVVKEGRESRLVVYFDRIMQGIEVSIYPELLSPIHAVGTEEQNE TGFY
5	MESFLNWYLIYTKVKKEDYLEQLLTEAGLEVLNPKIKKTKTVRNKKKEVIDPLFPCYLFV KADLNVHLRIISYTQGIRRLVGGSNPTIVPIEIIDTIKSRMVDGFIDTKSEEFKKGDTIL IKDGPFKDFVGIFQEELDSKGRVSILLKTLALQPRITVDKDMIEKLHN
6	MSKKWYAIQSKPNKEQALCEQFQSRGIEVFYPQIRVNPVNPRARKIRPYFPGYLFVHVDL DEVGLSVIRWIPFARGVVSFSNEPASVPDNLIEAIRRRVDEVNRAGGELLETLKPGEPVL IQEGPFAGYEAIFDVRLSGKERVRVLIQLLSQRYIPVEMQVGSLKPLKTKNKDKPHPL
7	MSEQQKYWFAARTRDKQEFAIRDSLEKLKTELDLNYYLPTQFVIRQLKYRRKRVEVPVIK NLIFIQATKQDACDISNKYNIQLFYMKDLLTRAMLIVPDKQMQDFIFVMDLDPNGVSFDN DHLSVGSRVQVVKGDFCGVEGELASEANKTYVVIRIAGVLSASVKVPKSYLRVI
8	MARKWYAVRTYSGHENRVKKFIENEIAEGKLKDKIFNVLVPTEKVTVVKEGRKRSRVKAF FPGYILIEAEMDDEVKNFIRSVPSVVSFVGPKGNPVPLREDEVERFVGKGEEVGEVERVD VPFRVGDSVKVIDGPFADFSGIVQEVNSGKMKLKVMINIFGRKTPVELDFAQVEIEK
9	MSEAPKKRWYVVQAFSGFEGRVATSLREHIKLHNMEDLFGEVMVPTEEVVEIRGGQRRKS ERKFFPGYVLVQMVMNDASWHLVRSVPRVMGFIGGTSDRPAPISDKEVDAIMNRLQQVGD KPRPKTLFEPGEMVRVNDGPFADFNGVVEEVDYEKSRLKVSVSIFGRATPVELDFSQVEK A
10	MRMDEGLSRSGGDRVAKQWYIVHTYSGFEHRVKAALQERIKAAGKEEYFGQILVPTEKVV ELVKGERKSSSRKFYPGYIVVEMELNDETWHLVRHTPKVTGFIGSQERPIPLSEEEANAI IQQMEEGIQKPRPKYQFEKGEEVRVVDGPFASFNGVVEQVIPEKGKVRVLVTIFGRSTPV ELDFVQIQRL

Data S1. Annotations and predictions of all sequences identified in the NusG superfamily.

Data S2. Additional annotations and predictions of archaeal and eukaryotic sequences used to determine the tree in **Figure 4**.