Many dissimilar NusG protein domains switch between α-helix and β-sheet folds

Lauren L. Porter1,2✉, Allen K. Kim1, Swechha Rimal1,2, Loren L. Looger3, Ananya Majumdar4, Brett D. Mensh3, Mary R. Starich2 & Marie-Paule Strub2

Folded proteins are assumed to be built upon fixed scaffolds of secondary structure, α-helices and β-sheets. Experimentally determined structures of >58,000 non-redundant proteins support this assumption, though it has recently been challenged by ~100 fold-switching proteins. Though ostensibly rare, these proteins raise the question of how many uncharacterized proteins have shapeshifting—rather than fixed—secondary structures. Here, we use a comparative sequence-based approach to predict fold switching in the universally conserved NusG transcription factor family, one member of which has a 50-residue regulatory subunit experimentally shown to switch between α-helical and β-sheet folds. Our approach predicts that 24% of sequences in this family undergo similar α-helix ⇌ β-sheet transitions. While these predictions cannot be reproduced by other state-of-the-art computational methods, they are confirmed by circular dichroism and nuclear magnetic resonance spectroscopy for 10 out of 10 sequence-diverse variants. This work suggests that fold switching may be a pervasive mechanism of transcriptional regulation in all kingdoms of life.
or over 60 years, biological science has been heavily influenced by the protein folding paradigm, which asserts that a protein assumes one fold specified by its amino acid sequence. Fold-switching proteins challenge this paradigm by remodeling their secondary and tertiary structures and changing their functions in response to cellular stimuli. These proteins regulate diverse biological processes and are associated with human diseases such as cancer, malaria, and COVID-19. Nevertheless, the ostensible rarity of fold switching leaves open the question of whether it is a widespread molecular mechanism or a rare exception to the well-established rule. A major barrier to assessing the natural abundance of fold-switching proteins has been a lack of predictive methods to identify them. Whereas computational methods for rapid and accurate prediction of secondary and tertiary structure for single-fold proteins have been established, methods to simply classify fold switchers have lagged. This comparative lack of progress arises from the small number of experimentally observed fold switchers (<100), hampering the discovery of generalizable characteristics that distinguish them from single folders. As a result, essentially all naturally occurring fold switchers have been discovered by chance.

Previously, we developed a sequence-based approach to predict protein fold switching. This approach is based on the observation that the secondary structure of a protein domain or subdomain can change dramatically depending on its context. Accordingly, the secondary structure prediction of a fold-switching sequence can change depending on whether it is queried within part of a larger sequence or in isolation. Context-dependent secondary structure is rarely captured by conventional approaches, which tend to predict protein structure using a full amino acid sequence only. (Fig. 1b) or sub-sequences (“crops”) significantly longer than fold-switching regions. This problem can be circumvented by comparing secondary structure predictions of whole fold-switching sequences with isolated short (25-40-residues) fragments that could potentially switch folds. Predictions that shift from β-sheet to α-helix—or vice versa—by changing sequence context indicate fold switching with high statistical significance. Secondary structures were predicted using JPred4, a single-hidden-layer neural network trained on 1000 sequence-diverse proteins with solved structures. Previous work showed that JPred4 rarely mistakes β-helices for β-sheets—or vice versa—in single-fold proteins. Furthermore, it predicts fold switching more accurately than other secondary structure predictors because (1) it uses a curated database of non-redundant sequences and (2) it relies primarily on hidden Markov Models (HMMs) rather than position-specific scoring matrices (PSSMs). HMMs are more sensitive than PSSMs because they assume that insertion and deletion probabilities vary with sequence position and calculate insertion and deletion penalties from input sequence alignments rather than using ad hoc parameters. For instance, this sensitivity allowed JPred4 to predict dramatic changes in secondary structure resulting from a single amino acid substitution.

Fold-switching has been shown to occur in the NusG protein superfamily, which comprises both single-fold and fold-switching proteins. NusGs are the only family of transcriptional regulators known to be conserved from bacteria to humans. 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 often promote transcription elongation by reducing RNAP pausing. NusG homologs from other kingdoms of life, such as DSIF in humans and SpT5 in archaea and yeast, are also called NusGs in this paper. Specialized NusGs (NusGSPs), which also exist in all kingdoms of life, promote transcription elongation at specific operons only. Furthermore, some NusGs and NusGSPs couple transcription with other biological processes such as translation, RNA silencing, and chromatin modification. Atomic-level structures of NusGs from several organisms have been determined. Bacterial NusGs share a two-domain architecture with a NusG N-terminal (NGN) domain that binds RNAP, and a C-terminal Kyprides, Ouzounis, Woese (KOW) domain, which assumes a β-roll fold. The structure of Escherichia coli NusG is shown in Fig. 1a. The only NusGSP with an experimentally determined full-length structure is E. coli RfaH (Fig. 1a), whose sequence is 19% identical and 37% similar to that of E. coli NusG. While the N-terminal domain of E. coli RfaH maintains the NGN fold and RNAP-binding activity of its housekeeping NusG homologs, its C-terminal domain (CTD) 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. Thus, RfaH’s fold-switching CTD serves two purposes: (1) to regulate the N-terminal domain (NTD) so that it associates with RNAP exclusively at ops sites and (2) to foster efficient translation of transcripts produced by RNAP when bound to its NTD.

In this work, we focus on bacterial two-domain NusGs and hypothesize that the CTDs of fold-switching NusGs, such as RfaH, are predisposed to fold into both α-helical and β-sheet structures while single-fold NusGs are predisposed to fold into β-sheets only. Accordingly, RfaH’s CTD folds into an α-helical hairpin when expressed with its N-terminal NGN domain but into a β-roll when expressed in isolation. Thus, our approach compares the predicted secondary structures of both the full-length amino acid sequence (N-terminal NGN domain+CTD) and the isolated (cropped) C-terminal domain. CTDs with predicted β-sheet secondary structures in both full-length and cropped sequences are expected to be single folders with constant β-sheet propensities. By contrast, CTDs with regions whose predicted secondary structures change from β-sheet to α-helix when their sequences change from full-length to cropped are expected to switch folds. Applying this approach to all ~15,000 sequences in the NusG superfamily, our approach predicted that 24% of NusG-like sequences switch between α-helix and β-sheet folds, a proportion significantly larger than the 0.5-4% predicted previously.

Results

Pervasive fold switching predicted in the NusG superfamily. Our approach was tested on 15,516 nonredundant NusG/NusGSP sequences (Methods, Supplementary Data 1). Consistent with other methods (Fig. 1b), it predicted that 95% of CTDs would assume β-sheet folds when full-length sequences were used as inputs. By contrast, 24% of cropped CTD sequences (>3600) were predicted to have substantial α-helical content (Methods), suggesting that they switch folds. These prediction differences likely arise from the multiple sequence alignments (MSAs) used to generate predictions (Fig. 1c). N_q values, which quantify MSA depth and diversity, were ~3800 larger, on average, for PSI-BLAST-generated MSAs from full-length input sequences than from their cropped counterparts (Fig. S1, Methods). Thus, full-length alignments tend to be >3X deeper than cropped. As evidenced by the higher level of predicted β-sheet, these deeper, more diverse alignments seem to reflect properties of the NusG superfamily, whose CTDs can presumably fold into β-roll structures regardless of whether they switch folds. Conversely, shallower CTD alignments seem to reflect properties of NusG
subfamilies, some of whose members have CTDs with helical propensities, such as *E. coli* RfaH, while others maintain β-sheet propensities, such as *E. coli* NusG. To estimate the false-negative and false-positive rates of these predictions, we exploited known operon structures of NusG and several specialized homologs as an orthogonal method to annotate sequences as NusGs or NusG SPs. We mapped the sequences used for prediction to solved bacterial genomes (Methods) and analyzed each sequence’s local genomic environment for signatures of co-regulated genes. Of the 15,195 total sequences, 5175 mapped to contexts consistent with housekeeping NusG function. Only 26 of these were predicted to switch folds, suggesting a false-positive rate of 0.5% for fold-switch predictions. Performing a similar calculation in 849 previously identified RfaHs (Supplementary Data 1), 31 were predicted single folders. These results suggest that fold switching is widely conserved among RfaHs, which, if correct, indicates a false positive rate of 4% (31/849). Full-length *Vibrio cholerae* RfaH, whose sequence is 44% identical to *E. coli* RfaH, was characterized by NMR and found to have a helical CTD in a recent preprint. This result further indicates that fold switching is conserved among RfaHs. Of the remaining 8661 sequences with high-confidence predictions (Methods) – encompassing several NusG SP clades – 31% were predicted to switch folds.
Experimental validation of fold-switch predictions. A representative group of variants with dissimilar sequences was selected for experimental validation. First, all NusG superfamily sequences were clustered and plotted on a force-directed graph, hereafter called NusG sequence space (Fig. 2a, Supplementary Fig. 2, Supplementary Data 1). The map of this space, in which clusters with higher sequence similarity are grouped closer in space, revealed that some putative fold-switching/single-folding nodes are clustered together within sequence space (upper/lower groups of interconnected nodes), while other regions had mixed predictions (left/right groups of interconnected nodes). Sixteen candidates selected for experimental validation came from distinct nodes, had diverse genomic annotations, and originated from different bacterial phyla (Supplementary Fig. 3, Supplementary Tables 2, 3). Of these 16 candidates, 10 could be expressed and purified (Supplementary Table 1).

Circular dichroism (CD) spectra of 10 full-length variants were collected. We expected the spectra of fold switches to have more helical content than single folders because their CTDs have completely different structures (RfaH: all α-helix ground state, NusG: all β-sheet ground state), while the secondary structure compositions of their single-folding 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 (Fig. 1a)38,39. As expected, their CD spectra were quite different (Supplementary Fig. 4a): *E. coli* RfaH had a substantially higher α-helix/β-strand ratio (1.1) than *E. coli* NusG (0.57) – consistent with solved structures (Fig. 2b, variants #3 and #9).

All remaining predictions were also consistent with the CD spectra of their corresponding variants (Fig. 2b, Supplementary Table 2). Specifically, five predicted fold switches had RfaH-like CD spectra with minima at 208 nm, a characteristic feature of helical folds that suggests ground-state helical bundle conformations in two RfaHs (variants #2, #6), a LoaP (variant #1), an annotated NusG (variant #4), and an annotated “NGN domain-containing protein” (variant #5). Interestingly, all five of these variants had essentially as much predicted helical content as the reference fold switch, *E. coli* RfaH (α-helix/β-strand ratio ≥1.1), further suggesting ground-state helical CTDs. Additionally, the remaining three predicted single folders had NusG-like CD spectra that lacked minima at 208 nm: 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 work25 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 nine isolated CTDs: six from putative fold switches and three from putative single folders; the tenth (Variant 7 CTD) was degraded during expression on two independent occasions. All spectra had low helical content and high β-sheet content (Supplementary Fig. 4b), suggesting that the CTDs of all six predicted fold switches can assume α-helical hairpin folds in their full-length forms and β-roll folds when expressed in isolation.

CD can potentially mislead since it shows aggregate, rather than residue-specific, protein properties. Thus, it is possible that the higher helical content observed in Variants 1-6 could result from their NTDs rather than their CTDs. Though unlikely, since the NGN fold of the NTD is conserved from bacteria to humans23, we investigated this possibility for two variants at higher resolution using nuclear magnetic resonance (NMR) spectroscopy – which assigns residue-specific structure. Previous work25 has shown that the isolated CTD of RfaH, which folds into a β-roll, has a significantly different 2D 1H,15N heteronuclear single quantum coherence (HMQC) spectrum than full-length RfaH, whose CTD folds into an α-helical hairpin. Thus, we conducted similar experiments on one single-folding variant (#8) and one putative fold switcher (Variant #5). The backbone amide resonances of the full-length and CTD forms of Variant #8 were 98% superimposable, whereas only 12% of peaks from the full-length and CTD forms of Variant #5 overlapped (Fig. 2d). 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 can suggest a fold switch, though large shifts can also result from changes in CTD:NTD interactions without a significant conformational shift29. Subsequently, assigned backbone amide resonances were used to characterize the secondary structures of Variant CTDs at higher resolution using TALOS-N40 (Methods, Fig. 2d, Supplementary Fig. 4c, Supplementary Table 3). Both isolated CTDs had secondary structures consistent with the β-roll fold. Combining this result with the 98% peak overlap between full-length Variant #8 and its CTD (Fig. 2d) indicates that Variant #8’s CTD maintains a β-roll fold. Alternatively, the TALOS-N secondary structure predictions calculated from chemical shift assignments of the full-length Variant #5 CTD indicate that it is largely helical (Fig. 2d), demonstrating that it switches folds.

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

1 Some, but not all, NusGSPs besides RfaH probably switch folds. Specifically, full-length LoaP (variant #1), which regulates the expression of antibiotic gene clusters41, had an RfaH-like CD spectrum, whereas full-length UpbY, a capsular polysaccharide transcription antiterminator from
The fold-switching mechanism appears to be conserved among annotated RfaHs with low sequence identity (≤32%, variant #2, #3, and #6), a possibility proposed previously, though without experimental validation. Also, “NGN domain-containing protein” variant #5 is genomically inconsistent with being a NusG and is likely an RfaH.

Some annotated NusGs have RfaH-like CD spectra (variant #2, #3, and #6), a possibility proposed previously, among annotated RfaHs with low sequence identity (≤32%, variant #4), likely the result of incorrect annotation. Indeed, the genomic environment of variant #4 (Methods) suggests that it is a UpxY, not a NusG. Genomic environment of variant #4 (Methods) suggests that it is a UpxY, not a NusG.

Variant #5 is genomically inconsistent with being a NusG and is likely an RfaH. The 1H-15N HSQCs of full-length and CTD variants of a putative single-folder (Variant #8) are nearly superimposable. By contrast, the HSQCs of full-length and CTD variants of a putative fold switcher (Variant #5) differ significantly.

Other predictive methods do not capture the helical ground state of fold-switching variants. To benchmark the performance of our secondary-structure-based approach, we assessed whether machine learning and template-based methods could also distinguish between fold switchers and single folders in the NusG superfamily. Specifically, we tested AlphaFold2, Robetta, EVCouplings, and Phyre2 on variants #1-6, whose CD spectra were all RfaH-like, suggesting that their CTDs assume ground-state helical folds. All methods predicted only one CTD conformation per variant – almost all of which were β-sheet (Supplementary Fig. 5), except for AlphaFold2’s predictions of E. coli RfaH (variant #3), whose experimentally determined structure was in its training set, and variant #6, whose sequence is nearest and best connected with E. coli RfaH in Sequence Space (Fig. 2a).

Predicted amino acid contacts from Robetta and EVCouplings corresponded with the NusG-like contacts generated from these sequences that our approach predicted to switch folds (Methods). Source data are provided as a Data File.
the GREMLIN algorithm—was responsible for the discovery of alternative contacts. These results again indicate that shallower alignments—such as the JPred-filtered one ($N_{\text{eff}} = 834$)—reflect folding properties of the RfaH subfamily while deeper alignments ($N_{\text{eff}} = 9633$ and 17,245 for Robetta and EVCouplings, respectively), reflect folding properties of the NusG superfamily, whose members largely do not switch folds.

This analysis of putative fold-switching CTDs indicates evolutionary coupling of residue-residue contacts unique to two distinct folds. For the $\alpha$-helical fold, six intrahelical hydrophobic

**Fig. 3 Fold-switching sequences have conserved amino acid contacts from both folds.** Predicted amino acid contacts from fold-switching sequences (dark gray circles) correspond to both the $\beta$-roll fold (PDB ID: 2LCL, red circles) and the $\alpha$-helical hairpin fold (PDB ID: 5OND, chain A, teal circles). Couplings that do not correspond to experimentally observed contacts are shown as light circles. Categories of amino acid contacts from both folds use the alphabetically labeled contacts in the plot above them. Source data are provided as a Source Data file.
contacts and one set each of interhelical contacts, strand-helix contacts, and helix-capping contacts were observed (Fig. 3). Overall, 96% of interhelical contacts were hydrophobic, 94% of helix-capping residues could potentially form an i→4→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 were both hydrophobic. The remaining contacts gave more mixed results.

**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 (teal, 20.4%) than single folders (red, 40.5%). **b** Box-and-whiskers plots of pairwise sequence identities of fold-switching and single-folding CTDs of variants 1-10 in Fig. 2. The distributions of each teal (fold-switching)/red (single-fold) box were derived from n = 5/3 independent pairwise identities; each box bounds the interquartile range (IQR) of the data (first quartile, Q1 through third quartile, Q3); medians of each distribution are shown in white; lower whisker is the lowest datum above Q1-1.5*IQR; upper whisker is the highest datum below Q3 + 1.5*IQR. These distributions are consistent with the violin plots in panel (a). **c** Fold-switching CTDs are predicted in many bacterial phyla (blue background) 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 (green background) and Eukaryota (yellow background). Fold-switching/single-folding predictions are represented by teal/red colorings; predictions in branches with fewer than 10 sequences are gray. Source data are provided as a Source Data file.
Fold-switching CTDs are diverse in sequence, function, and taxonomy. It might be reasonable to expect fold-switching CTD 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%)\(^2\). The opposite is true. Sequences of putative fold-switching CTDs are substantially more heterogeneous (20.4% mean/19.4% median sequence identity) than sequences of predicted single folders (40.5% mean/42.5% median sequence identity, Fig. 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, Fig. 4b). Furthermore, fold-switching CTDs were predicted in most bacterial phyla, and many were predicted in archaea and eukaryotes as well (Fig. 4c, Supplementary Data 2). 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.

**Discussion**

Why might the sequence diversity of fold-switching CTDs exceed that of single folders? Functional diversity is one likely explanation\(^6\). Previous work has shown that NusG\(^{50}\)s drive the expression of diverse molecules from antibiotics to toxins\(^24\). Our approach suggests 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. Fold-switching NusG\(^{50}\)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 imminent desiccation. Supporting this possibility, NusG\(^{50}\) usually drive operons controlling rapid response to changing environmental conditions such as macrolide antibiotic production\(^41\), antibiotic-resistance gene expression\(^24\), virulence activation\(^47\), and biofilm formation\(^48\).

Our approach 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 containing 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\(^31\). Computational approaches that account for conformational variability and dynamics, a weakness in even the best predictors of protein structure\(^3\), 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\(^49\), and it could inform better design of fold switchers, a field that has seen limited success\(^50-52\).

Our results indicate that fold switching is a pervasive, evolutionarily conserved mechanism. Specifically, we predicted that 24% of the sequences within a ubiquitous protein family switch folds and observed coevolution of residue-residue contacts unique to both folds. This sequence-diverse dual-fold conservation challenges the protein folding paradigm and indicates that foundational principles of protein structure prediction may need to be revisited.

This work has two major limitations. Firstly, the level of error in our predictions is unknown. Due to the limited number of known fold-switching proteins, robust error rates of JPred4 as a fold-switch predictor cannot be determined. Although our experimental results suggest that the approach is accurate in at least ten cases tested, it is uncertain how well it performs in the full NusG superfamily (~15,000 proteins) or on proteins in general. Our orthogonal computational analysis indicates that the predictions capture single-folding NusGs 99.5% percent of the time. It is less clear how accurately they capture fold switching in NusG\(^{50}\)s, since some do not switch folds (e.g., Variant #7) and others do (Variants 1–6). Thus, additional work is needed to determine error rates and sources of error from this approach. Secondly, CD does not provide residue-specific structural information. Thus, it is possible that helical character arising outside of NusG CTDs could lead to the RfaH-like CD spectra observed in Variants 1–6. This possibility seems unlikely, however, given that all 6 variants are two-domain proteins whose N-terminal NGN domains are highly conserved\(^23\). Furthermore, Variants #3 and #5 have been shown by NMR previously\(^2\) or here to assume two folds.

The success of our method in the NusG superfamily suggests that it may have enough predictive power to identify fold switching in protein families where only single folders have been observed to date. 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.

**Methods**

**Identification of NusG-like sequences.** NusG-like sequences were identified from the October 2019 Uniprot\(^{90}\) database\(^3\) using an iterative BLAST\(^3\) approach. Specifically, the *E. coli* RfaH sequence (Uniprot ID Q0TAL4) was BLASTed against the database. All hits with a maximum e-value of 10\(^{-7}\) were aligned using Clustal Omega, 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\(^55\). Sequence identity between proteins in each cluster was ≥78%. Randomly selected sequences from the 25 largest clusters were then individually BLASTed against the Uniprot\(^{90}\) 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\(^24\). 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 P0A9FG0) 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 per CTDs long enough to make these predictions totaled 15,195 (Supplementary Data 1), 98% of all NusG-like sequences identified.

**JPred4 predictions.** JPred4\(^2\) predictions were carried out as in\(^1\), as follows. PSI-BLAST\(^3\) searches were run all 50-residue CTD sequences using two databases: the JPred database (http://www.compsi.dundee.ac.uk/jpred/about_RETRET_Nev231_details.shtml) from 2014 and the Uniprot\(^{90}\) database from January 2021. The resulting sequences were aligned with MView\(^4\) and inputted into HMMER\(^2\) 3.2.2 to generate a Hidden Markov Model (HMM). The resulting HMM was converted to GCG using hmmconv and converted to JPred4 input using the activation
Expression and purification of variants 1-16. Genes encoding all variants were ordered from IDT as gBlocks; all were codon optimized for E. coli. Except for the gene encoding variant 1, these genes were digested with HindIII and EcoRI and incorporated into the pPAP7 vector (Bio-Rad) with an N-terminal 6-His tag cloned using a Q5 mutagenesis kit (New England Biolabs); we call this modified vector HisPAP7. In further detail, hisPAP7 was also digested with HindIII and EcoRI. Digested plasmid and digested genes were individually purified with a QIAquick PCR Purification kit (Qiagen). Then, T7 DNA Ligase and T7 DNA Ligase buffer (New England Biolabs) diluted to 1X final concentration. Reactions were incubated at room temperature for 1 hour, and 5 μL of each reaction was transformed directly into E. coli DH5α cells (New England Biolabs), and plated on Luria Broth (LB) agar plates with 100 μg/mL ampicillin overnight. Two colonies from each plate were picked individually and grown overnight in 3 mL LB with 100 μg/mL ampicillin at 37°C, shaking at 225 RPM. Plasmids were purified using the Qiagen Spin Miniprep Kit (Qiagen), and the genetic sequences of each variant were confirmed by Sanger sequencing (Psmagen). Plasmids with confirmed genetic sequences were transformed into E. coli BL21-DE3 cells (New England Biolabs), grown in LB at 37°C to an OD600 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. The gene encoding variant #7 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 10,000xg 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 8.0) and frozen at ~80°C for later purification. Sequencing of all variants was verified by Psmagen.

Thawed cell pellets were resuspended in 25 ml lysis buffer per 1 L of culture grown. 100 mg of DNAseI, 5 mM CaCl2, 5 mM MgSO4 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 a French pressure homogenizer (Avestin). The homogenized lysate was centrifuged for 45 minutes at 40,000xg at 4°C, and its soluble fraction was loaded immediately onto either a 1 mL Ni column (GE HiTrap HP) or an Econo-Pac (Bio-Rad) gravity column with 0.5-1 mL IMAC Ni Resin (Bio-Rad). Soluble lysis was stored on ice while loading at 1 mL/minute through the sample pump of a room-temperature AKTA Avant onto a 1 mL HisTrap column or 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 AKTA 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 2 M, all in 100 mM phosphate, pH 7.4-8.0.

Nickel-purified samples were then loaded onto 1- or 5-mL ProteinX eXact columns (Bio-Rad), washed twice with one column-volume of 2 M 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), the buffer was exchanged to 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 AKTA 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 2 M, all in 100 mM phosphate, pH 7.4-8.0.

Variant CTDs. Full-length variants were shortened using Q5 mutagenesis (New England Biolabs; oligonucleotide sequences are in Supplementary Table 4). Their sequences were confirmed by Sanger sequencing (Psmagen) and are reported in Supplementary Table 3. TS or TSW tags were added to most constructs (but not V1-V6 CTDs) to speed up their cleavage. The XaaX³ column and to improve concentration measurements using absorbance at 280 nm. All variants were expressed and purified as were variants 1-16, using expression temperatures of 20°C for Variants 2, 5, 8, and 9 and 16°C for the rest. We attempted to purify Variant 7 CTD twice, but it showed signs of degradation during expression in both instances.

Circular dichroism (CD) spectroscopy. CD spectra of all samples were measured within 1-2 days of purification; they were stored at 4°C until then. All CD spectra were collected on Chirascan spectrometers (Applied Photophysics) in 1 mm quartz cuvettes (Hellma) in 100 mM phosphate, pH 7.4. Protein concentrations ranged...
from 8 to 12 μM, and scan numbers ranged from 5 to 10, collected at 1 nm/s with a 1 nm step size. Scans were averaged, and averaged baselines of buffer-blank 1 mm cuvettes were subtracted from the spectra. The resulting spectra were converted to units of Molar Residue Ellipticity \([\theta]_{222}\) using Eq. (3):

\[
[\theta]_{222} = \frac{\theta}{c} 
\]

where \(\theta\) is the ellipticity measured by the instrument, \(c\) is the extinction coefficient determined by Exasy Proptramat\(^\text{68}\), \(L\) is the path length of the cuvette, \(N\) is the number of amino acids, and \(A\) is the absorbance measured by a Nanodrop One (Thermo Scientific). Absorbances were measured at 280 nm for all full-length constructs (Supplementary Table 2) as well as the CTDs of Variants 2, 6, 7, and 9, to which a 10% D2O was added to the N-terminal groups (Supplementary Table 3); sequence-based extinction coefficients of these variants were calculated using Exasy Proptramat\(^\text{64}\) (https://web.expasy.org/proptramat/). Absorbances of Variants 1, 4, 6, and 10 were measured at 205 nm, with extinction coefficients calculated from https://spinn.midd.nh.gov/cloneSoftware/A203.html\(^\text{85}\). Concentrations of each of the CTDs of Variants 3, 5, and 8 were determined using the Bradford Assay against a Bovine Serum Albumin (New England Biolabs) baseline measured with concentrations of 0.1, 0.25, 0.5, 1.0, and 2.0 mg/mL. Concentrations were converted to molarities based on molecular weights calculated using Exasy Proptramat\(^\text{64}\). Resulting spectra were entered into the BestSel\(^\text{66}\) webserver (https://bestsel.elte.hu/index.php), so that their ratio of helix (helix + strand: strand (parallel + antiparallel)) could be computed. Ratios were calculated for two wavelength ranges (195–250 nm and 200–250 nm) and averaged (Fig. 2b, Source Data).

Expression and purification of NMR samples. Based on the protocols in\(^\text{67–69}\), E. coli BL21 DE3 cells (New England Biolabs) expressing all isotypically labeled samples were grown in LB to an OD\(_{600}\) of 0.6 and pelleted at 5000×g for 30 minutes at 4 °C. The pellets were resuspended in 1X M9 at half of the initial culture volume and pelleted at 10,000×g for 30 minutes at 4 °C. Pellets were then resuspended at ¼ final culture initial volume in 2X M9, pH 7.0–7.1, 1 mM MgSO\(_4\), 0.1 mM CaCl\(_2\), with 1 g NH\(_4\)Cl/L, and 4 g of either unlabeled or 13\(^\text{C}\)-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 10,000×g for 10 minutes at 4 °C.

All labeled variants were purified by FPLC (AKTA Avant 25) using the same methods as variants 1–10 above in 5 mM HisTrap HP columns (Cytiva) and 5 mM Proflint eXact columns (BioRad).

\(^{1}\text{H}-^{15}\text{N}\) HSQCs of Variants \#5 and \#8. All spectra were collected on Bruker Avance II 600 MHz spectrometers equipped with z-gradient cryoprobes and processed with NMRpipe\(^\text{70}\). Assignments of Variant \#5 CTDs and Variant \#8 CTDs were determined from the upper triangular portions of each matrix, excluding positions where the CTDs of Variants 3, 5, and 8 were determined using the Bradford Assay against a Bovine Serum Albumin (New England Biolabs) baseline measured with concentrations of 0.1, 0.25, 0.5, 1.0, and 2.0 mg/mL. Concentrations were converted to molarities based on molecular weights calculated using Exasy Proptramat\(^\text{64}\). Resulting spectra were entered into the BestSel\(^\text{66}\) webserver (https://bestsel.elte.hu/index.php), so that their ratio of helix (helix + distorted helix):strand (parallel + antiparallel) ratio could be computed. Ratios were calculated for two wavelength ranges (195–250 nm and 200–250 nm) and averaged (Fig. 2b, Source Data).

Phylogenetic tree. The tree in Fig. 4c was generated by downloading the Interactive Tree of Life\(^\text{77}\) (https://itol.embl.de/itol.cgi), loading it into FigTree\(^\text{78}\), and collapsing branches at the phyletic level, except for Proteobacteria, which were left at the class level because of recent phylogenetic work on prototype bacterial Rfa\(_H\)\(^\text{24}\). Bacterial species from each NusG sequence were obtained from their Uniprot accession codes used in Fig.1: 6ZTJ [https://doi.org/10.2210/pdb6ZTJ/pdb] (E. coli RfaH). Protein sequences were precomputed by comparing the distances and angles 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-H angles ≥90° and helix caps if the distance between sidechain donor/acceptor 5° and C-N-C angles ≥90° were ≥90°. All distances and angles were calculated using LINUS\(^\text{74}\).

C TD sequences for GRELIN webserver (http://grelin.bakerlab.org/submit.php) analysis in Fig. 3 were obtained by clustering all PDB predictions by Affinity Propagation using the python scikit-learn module\(^\text{73}\) with 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, ––); 5, coiled-secondary structure discrepancies (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 GRELIN, 4 iterations of HHBNI\(^\text{72}\) were run on the initial alignment with e-values of 10–10. 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 portions of each matrix, excluding positions of identity, using nmpy\(^\text{76}\). Pairwise sequence identity matrices of the CTDs of the 10 variants were determined with Clustal Omega.

Data availability. Tabular data are provided in the Source Data File. Data recording all predictions are included in Supplementary Data 1 (bacteria and some archaea and eukaryotes) and Supplementary Data 2 (expanded predictions for archaea and eukaryotes). Additional data are available at https://github.com/ncc/reviews_sequence chemical. Chemical shift assignments were deposited in the Biomolecular Magnetic Resonance Bank (BMRB) with the following accession codes: 51429 [https://doi.org/10.1308/BMRB51429] (Full-length Variant 5 (CTD only)), 51428 [https://doi.org/10.1308/BMRB51428] (Variant 5 isolated CTD), 51431 [https://doi.org/10.1308/BMRB51431] [Variant 8 CTD]. PDB accession codes used in Fig. 1: 6ZTJ [https://doi.org/10.2210/pdb6ZTJ/pdb] (E. coli 70S-RNAP expression). Complex in NusG-coupled state, 38 nt intervening mRNA, chain CF, 50ND [https://doi.org/10.2210/pdb50ND/pdb] [RfaH from Enterobacteria coli in complex with ops DNA, chain A], and 6Q65 [https://doi.org/10.2210/pdb6Q65/pdb] (CryoEM
structure of the E. coli RNA polymerase elongation complex bound with RfaH, chain D). PDB accession codes used in Fig. 3: 5OND (https://doi.org/10.2210/pdb5OND/pdb) (RfaH from Escherichia coli in complex with rpsA DNA, chain A), 2LCL (https://doi.org/10.2210/pdb2LCL/pdb) (Solution Structure of RfaH carboxyterminal domain, chain A). Constructs for protein expression are available upon request. Source data are provided with this paper.

Code availability
Code used to generate the predictions reported in this manuscript can be found at: https://github.com/nch/sequence_space.

Received: 21 December 2021; Accepted: 17 June 2022; Published online: 01 July 2022

References
1. Anfinson, C. B. Principles that govern the folding of protein chains. Science 188, 220–230 (1975).
2. Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. Proc. Natl Acad. Sci. USA 115, 5968–5973 (2018).
3. Kim, A. K. & Porter, L. L. Functional and Regulatory Roles of Fold-Switching Proteins. Structure 29, 6–14 (2021).
4. Li, B. P. et al. CLIC1 Promotes the Progression of Gastric Cancer by recruiting to a hairpin structure of the non-template DNA strand. Elife 7, e36349 (2018).
5. Zuber, P. K. et al. Structural and thermodynamic analyses of the beta-to-alpha transformation in RfaH reveal principles of fold-switching proteins. bioRxiv https://doi.org/10.1102/2022.01.14.476317 (2022).
6. Goodman, J. R., Kluft, S., Zhang, C., Straight, P. & Winkler, W. C. LoaP is a broadly conserved anti-terminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. Nature Microbiol 2, 17003 (2017).
7. Wang, C. et al. Structural basis of transcription-translation coupling and collision in bacteria. Science 369, 1355–1359 (2020).
8. Wang, C. et al. Structural basis of transcription-translation coupling. Science 369, 1359–1365 (2020).
9. Chen, Y. & Bax, A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. Methods Mol. Biol. 1260, 17–32 (2015).
10. Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. Science 324, 203–207 (2009).
11. Leedh, J. A. & Welch, R. A. RfaH enhances elongation of Escherichia coli hlyCABD mRNA. J. Bacteriol. 178, 1850–1857 (1996).
12. Belouz, C. et al. The transcriptional antiterminator RfaH represses biofilm formation in Escherichia coli J. Bacteriol. 188, 1316–1331 (2006).
13. Monteith, W. B., Cohen, R. D., Smith, A. E., Guzman-Cisneros, E. & Pietak, G. J. Quinary structure modules protein stability in cells. Proc. Natl Acad. Sci. USA 112, 1739–1742 (2015).
14. Ambrogio, X. I. & Kuhlman, B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. J. Am. Chem. Soc. 128, 1154–1161 (2006).
15. Wei, K. Y. et al. Computational design of closely related proteins that adopt two well-defined but structurally divergent forms. Proc. Natl Acad. Sci. USA 117, 7208–7215 (2020).
16. UniProt, C. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142–D148 (2010).
17. Siwers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539 (2011).
18. Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
Brown, N. P., Leroy, C. & Sander, C. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14, 380–381 (1998).

Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39, W29–W37 (2011).

Cuff, J. A. & Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34, 508–519 (1999).

Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009).

Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl Acad. Sci. USA 110, 15674–15679 (2013).

Hagberg, A. A., Schult, D. A., and Swart, P. J. in Proceedings of the 7th Python in Science Conference. (ed Travis Vaught Giel Varoquaux, Jarrod Millman) 11-15.

Rost, B. Twilight zone of protein sequence alignments. Proteins Eng. 12, 85–94 (1999).

Ruan, B., Fisher, K. E., Alexander, P. A., Doroshko, V. & Bryan, P. N. Engineering subtilisin into a fluoride-triggered processing protease useful for one-step protein purification. Biochemistry 43, 14539–14546 (2004).

Gasteiger, E. et al. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31, 3784–3788 (2003).

Anthis, N. J. & Clore, G. M. Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. Protein Sci. 22, 851–858 (2013).

Micsonai, A. et al. BeStSel: webserver for secondary structure and fold prediction for protein CD spectra. Nucleic Acids Res https://doi.org/10.1093/nar/gkaa345 (2022).

Azatian, S. B., Kaur, N. & Latham, M. P. Increasing the buffering capacity of minimal media leads to higher protein yield. J. Biomol. NMR 73, 11–17 (2019).

Cai, M., Huang, Y., Yang, R., Craigie, R. & Clore, G. M. 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).

Marley, J., Lu, M. & Bracken, C. A method for efficient isotopic labeling of recombinant proteins. J. Biomol. NMR 20, 71–75 (2001).

Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR 6, 277–293 (1995).

Ying, J., Delaglio, F., Turchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. J. Biomol. NMR 68, 101–118 (2017).

Lee, W., Tonelli, M. & Marley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31, 1325–1327 (2015).

Kortemme, T., Morozov, A. V. & Baker, D. 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).

Srinivasan, R. & Rose, G. D. A physical basis for protein secondary structure. Proc. Natl Acad. Sci. USA 96, 14258–14263 (1999).

Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175 (2011).

Harris, C. R. et al. Array programming with NumPy. Nature 585, 357–362 (2020).

Letunic, I. & Bork, P. Interactive Tree Of Life (iTOl) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res https://doi.org/10.1093/nar/gka301 (2021).

FigTree v1.4 Molecular evolution, phylogenetics and epidemiology (2012).

Shen, W. & Ren, H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. J. Genet Genomics https://doi.org/10.1016/j.jgg.2021.03.006 (2021).

Eloe-Fadrosh, E. A. et al. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat. Commun. 7, 10476 (2016).

Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol 1, 16048 (2016).

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012).

Hunter, J. D. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95 (2007).

The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

Waskom, M. L. seaborn: statistical data visualization. J. Open Source Software 6, 3021 (2021).

Acknowledgements

L.L.P. thanks Marius Clore and Carolyn Ott for constructive discussions. We also thank George Rose, Laskin Swint-Kruse, Gisela Storz, Juan Bonifacino, Nico Tjandra, David Nynhuys, and Daniel Morris for helpful comments concerning the text and Drs. Juliette Lecomte and Christos Kougentakis for helping us to collect NMR spectra at the Johns Hopkins University Biomolecular NMR Center. This work utilized resources from the NHLBI Biophysics Core, the NHLBI Protein Expression Facility, and the NIH HPS Biowulf cluster (http://bpc.nih.gov). It was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health (LMM20211, L.L.P.) and Howard Hughes Medical Institute (L.L.P.).

Author contributions

Conceptualization: LLP, LLL. Methodology: LLP, LLL, AK, MS, AM, MPS. Software: LLP, LLL, AK. Investigation: LLP, LLL, AK, SR, MPS. Data Curation: LLL, LLP. Visualization: LLP, BDM, AK. Writing – original draft: LLP. Writing – review & editing: LLP, BDM, LLL, MS, AK. Supervision: LLP, LLL. Project administration: LLP. Funding acquisition: LLP, LLL.

Funding

Open Access funding provided by the National Institutes of Health (NIH).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31532-9.

Correspondence and requests for materials should be addressed to Lauren L. Porter.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article that is included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022.