A sequence-based approach for identifying protein fold switchers

Soumya Mishra[1], Loren L. Looger[1], and Lauren L. Porter[1,*]

Abstract

Although most proteins conform to the classical one-structure/one-function paradigm, an increasing number of proteins with dual structures and functions are emerging. These fold-switching proteins remodel their secondary structures in response to cellular stimuli, fostering multi-functionality and tight cellular control. Accurate predictions of fold-switching proteins could both suggest underlying mechanisms for uncharacterized biological processes and reveal potential drug targets. Previously, we developed a prediction method for fold-switching proteins based on secondary structure predictions and structure-based thermodynamic calculations. Given the large number of genomic sequences without homologous experimentally characterized structures, however, we sought to predict fold-switching proteins from their sequences alone. To do this, we leveraged state-of-the-art secondary structure predictions, which require only amino acid sequences but are not currently designed to identify structural duality in proteins. Thus, we hypothesized that incorrect and inconsistent secondary structure predictions could be good initial predictors of fold-switching proteins. We found that secondary structure predictions of fold-switching proteins with solved structures are indeed less accurate than secondary structure predictions of non-fold-switching proteins with solved structures. These inaccuracies result largely from the conformations of fold-switching proteins that are underrepresented in the Protein Data Bank (PDB), and, consequently, the training sets of secondary structure predictors. Given that secondary structure predictions are homology-based, we hypothesized that decontextualizing the inaccurately-predicted regions of fold-switching proteins could weaken the homology relationships between these regions and their overpopulated structural representatives. Thus, we reran secondary structure predictions on these regions in isolation and found that they were significantly more inconsistent than in regions of non-fold-switching

proteins. Thus, inconsistent secondary structure predictions can serve as a preliminary marker of fold switching. These findings have implications for genomics and the future development of secondary structure predictors.

Introduction

Most structurally characterized proteins perform one well-defined function supported by one dynamic scaffold of secondary structure [1, 2]. This observation, substantiated by over 100,000 atomic-level protein structures [3], has laid the basis for homology-based protein structure prediction algorithms. These algorithms are powerful enough to predict the structures of previously uncharacterized proteins with significant sequence similarity to a solved protein structure. If the sequence of a protein of interest falls below the threshold of significant similarity [4], sequence-based secondary structure predictions remain a viable option for model building. These lower resolution predictions can be used to develop putative three-dimensional protein structures that suggest their possible functional repertoire [5, 6].

Recent work has shown that a growing number of proteins can remodel their secondary structures in response to cellular stimuli, enabling radical functional changes and tight cellular control [7]. This phenomenon, called fold switching [8], can involve structural and functional transformations as drastic as an α-helical transcription factor morphing into a β-barrel translation factor [9]. Additionally, some fold-switching proteins change their cellular localization. For example, CLIC1 is a human protein that functions as both a cytosolic glutathione reductase [10] and a membrane-inserted chloride channel [11]. While full protein domains can switch folds, current experimental evidence suggests that it is more common for subdomains of larger proteins to switch folds while the remainder maintains a single native structure. We call the structurally changing subdomains "fold-switching regions" (FSRs) and the structurally unchanging remainders "non-fold-switching regions" (NFSRs).

Predicting the fold-switching ability of a given protein region can suggest a mechanism for its behavior *in vivo*, especially when predictions are combined with other forms of evidence indicating that the protein is multifunctional or has more than one cellular localization. Because fold-switching proteins challenge the assumption that globular proteins adopt one three-dimensional structure that performs one specific function, homology-based structure prediction methods are currently unable to predict multiple structures accessible to a given amino acid sequence. Furthermore, current *de novo* methods for predicting fold switching in proteins are also weak [12].

Previously, we successfully predicted FSRs by exploiting the failures of homology-based secondary structure predictors [7]. Specifically, we showed that discrepancies between predicted and experimentally determined secondary structures can indicate that a given protein switches folds. By coupling these discrepancies with a structure-based thermodynamic calculation [13], we were able to successfully predict fold switching in 13 proteins with one solved structure and experimental evidence for an alternative conformation. Thus, experimental discrepancies in secondary structure predictions could be an effective method for predicting whether a given amino acid sequence switches folds.

Here, we seek to determine if secondary structure discrepancies alone can suggest whether a given protein switches folds. This approach has the advantage of not requiring a solved protein structure with significant sequence similarity to the protein of interest. To lay the basis for this approach, we show that incorrect secondary structure predictions are significantly more common within FSRs than within NFSRs. We find that these incorrect predictions arise predominantly from one of the two conformations accessible to a given fold-switching protein. Inaccurately predicted conformers tend to be underrepresented in the Protein Data Bank (PDB), demonstrating that secondary structure predictions are influenced by structural bias within the PDB. We show that this bias can be leveraged to identify FSRs using secondary structure prediction discrepancies between FSR sequences in isolation and identical sequences within the context of their parent proteins. Thus,

secondary structure discrepancies can be used as a preliminary predictor of fold switching from amino acid sequence alone. This result has implications for the improvement of secondary structure predictors as well as identification of fold switching in proteins of biological interest.

Results

Secondary structure predictions of fold-switching proteins are consistently less accurate than those of non-fold switchers

First, we sought to determine whether secondary structure predictions of fold-switching protein regions (FPRs) are significantly less accurate than predictions of non-fold-switching protein regions (NFPRs) by comparing their secondary structure prediction accuracies. To do this, we ran three secondary structure predictors (JPred [14], PSIPred [15], and SPIDER2 [16]) on a set of 192 structures of fold-switching proteins [7] and compared the predicted secondary structure annotations of FPRs and NFPRs with the corresponding secondary structures determined by experiment. **Fig. 1a-c** shows the prediction accuracy distributions of both fold-switching and non-fold-switching protein regions. In all three distributions, the predictions of fold-switching regions were significantly less accurate than the predictions of non-fold-switching regions, with p-values of $8.6*10^{-309}$, $2.2*10^{-179}$, and $7.1*10^{-135}$, respectively (Kolmogorov-Smirnov Test).

To further test whether secondary structure predictions of FPRs are systematically less accurate than those of NFPRs, we generated secondary structure prediction accuracy distributions (**Fig. 1 d-f**) for randomly selected regions of proteins that are expected not to switch folds [7]. Upon comparing these histograms with those of fold-switching proteins, we again found significant differences in secondary structure prediction accuracies, with p-values of $1.2*10^{-264}$, $1.6*10^{-276}$, and $3.3*10^{-176}$, respectively (Kolmogorov-Smirnov Test). Together, these results indicate that secondary structure prediction accuracies of FPRs are significantly worse than the corresponding accuracies of NFPRs.

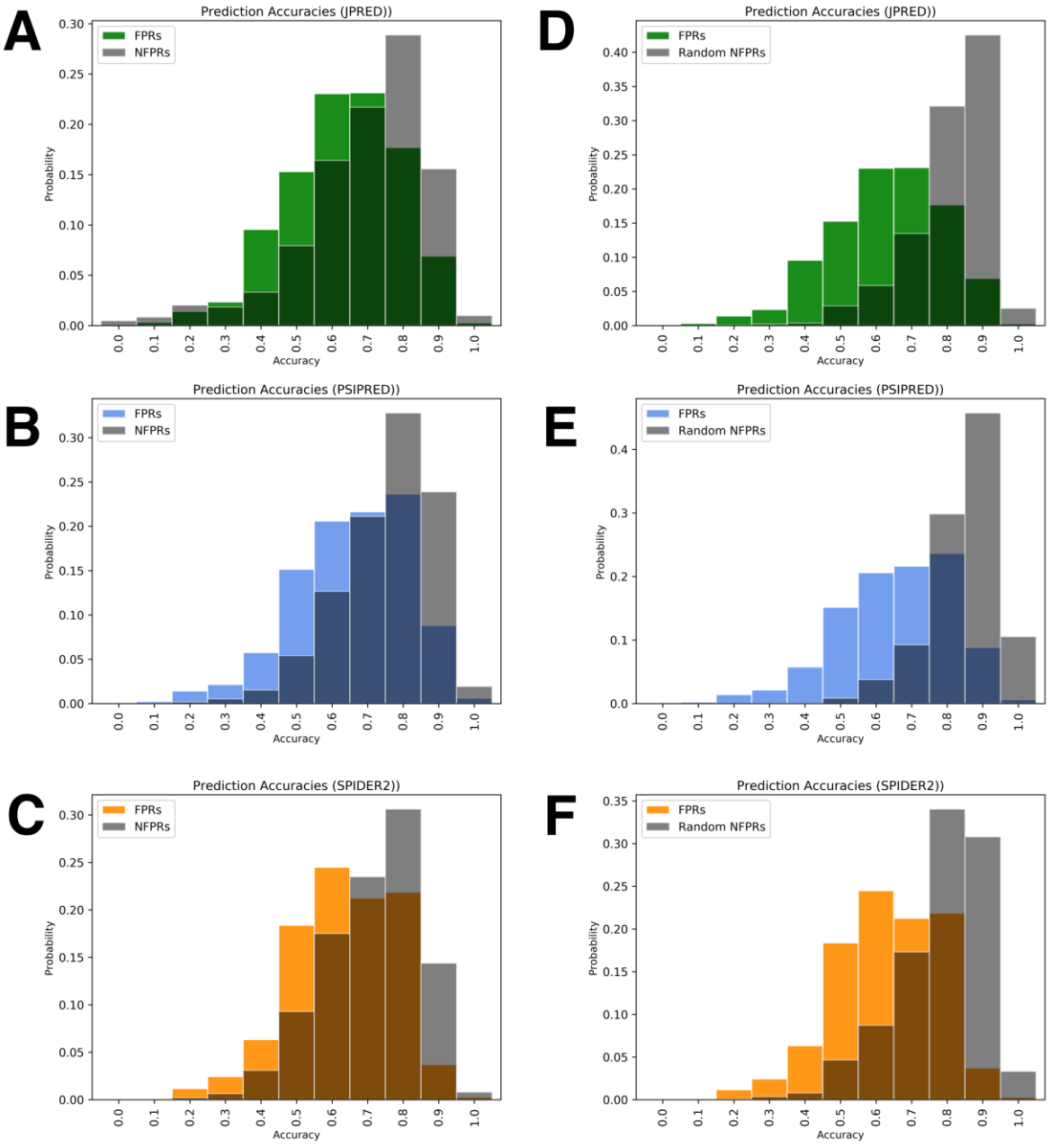
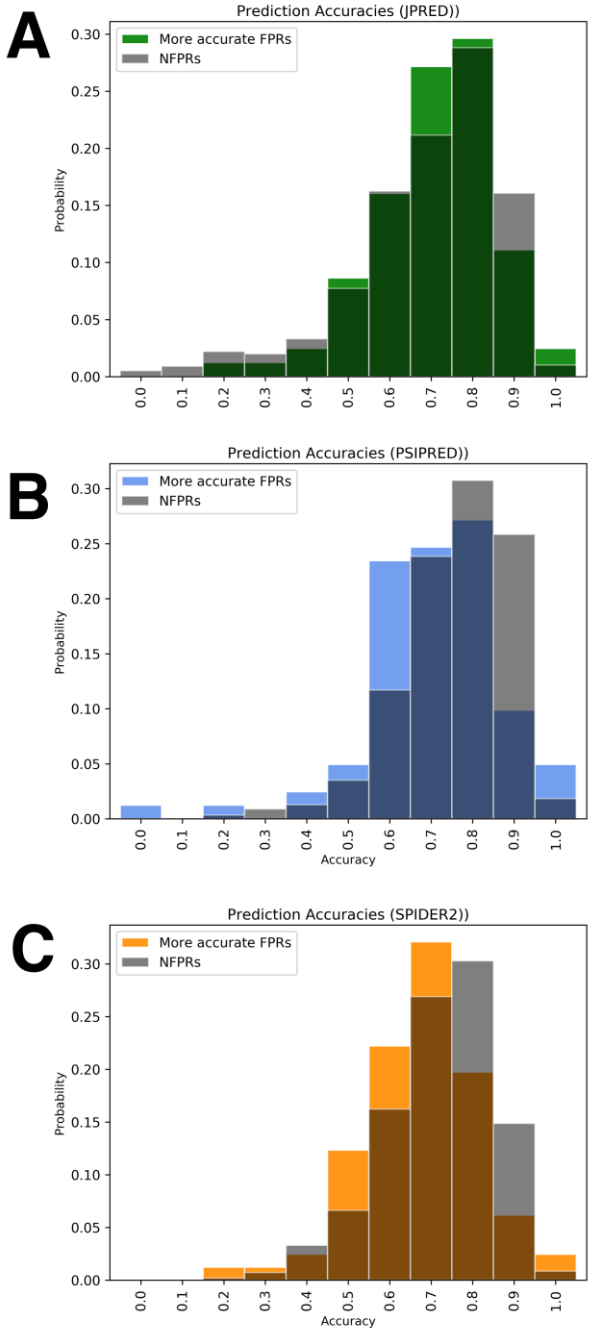


Figure 1. Secondary structure predictions of FPRs are consistently less accurate than those of NFPRs (a-c) and randomly-selected NFPRs (d-f). Histograms of fold-switching fragments are colored (green, JPRED; blue, PSIPRED; orange, SPIDER2), while comparisons of non-fold-switching fragments from corresponding predictors are gray.



We then sought to determine why secondary structure predictions of FPRs are significantly less accurate than NFPRs. We suspected that the conformational duality of FPRs contributed to inaccurate secondary structure predictions. JPRED, PSIPRED, and SPIDER2 are designed to predict only one secondary structure configuration for a given amino acid sequence. Thus, at best, they could predict only half of the FPR conformations correctly since FPRs have two distinct conformations. Alternatively, they could predict both FPR conformations equally inaccurately.

To investigate whether secondary structure predictions of one FPR conformation were, on average, significantly more accurate than those of the other, we remade our secondary structure prediction accuracy distributions for FPRs by including only the more accurate secondary structure prediction of the two conformations (**Fig. 2**). We found

Figure 2. Better FPR predictions are significantly more accurate and similar to NFPR distributions. Histograms of fold-switching fragments are colored (green, JPRED; blue, PSIPRED; orange, SPIDER2), while comparisons of non-fold-switching fragments from corresponding predictors are gray.

that the resulting FPR accuracy distributions were significantly more similar to those of NFPRs, with p-values of 0.88, 0.11, and 0.08, respectively (Kolmogorov-Smirnov test). Thus, JPRED, PSIPRED, and SPIDER2 tend to predict one FPR conformation with reasonable accuracy but predict the alternative poorly. The remaining discrepancies between FPR and NFPR inaccuracies result from SS predictor tendencies to predict the secondary structures of both FPR conformations inaccurately. This accounts for 27%, 32% and 40% of the predictions from JPRED, PSIPRED, and SPIDER2, respectively.

Conformational overrepresentation contributes to incorrect secondary structure predictions of FPRs

Upon noticing secondary structure prediction bias, we sought to determine its source. We hypothesized that FPRs with high frequencies within the PDB were more likely to be predicted correctly, while FPRs with fewer representative structures were more likely to be predicted incorrectly. This hypothesis was based on the observation that all three secondary structure predictors are trained on proteins with previously solved structures. Thus, training sets would more likely contain protein conformations that were highly represented within the PDB.

To test our hypothesis that structural bias within the PDB leads secondary structure predictors to accurately predict one FPR conformation but not the other, we determined the number of structures available for each FPR conformation (**Supp. Table 1**) and tested if FPRs with more accurate predictions tended to have more representative structures in the PDB. Our null hypothesis was that FPRs were equally likely to be predicted correctly, regardless of how many representative structures they had. By performing the Binomial Test, we disproved the null hypothesis (p-values of $7.7*10^{-4}$, $8.2*10^{-3}$, and $6.3*10^{-2}$, respectively), demonstrating that secondary structure predictions are biased towards predicting FPR conformations with frequent PDB representation and biased against predicting FPR conformations with infrequent PDB representation.

Secondary structure predictions of FSRs are context-dependent

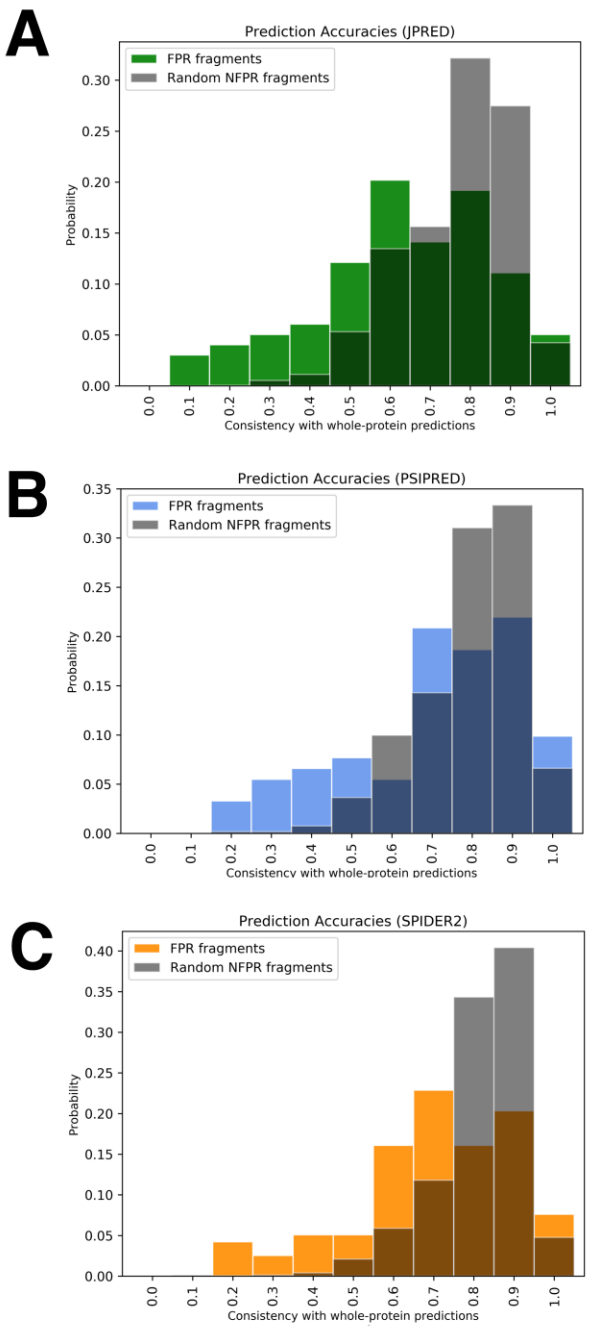


Figure 3. Consistencies between secondary structures of protein fragments in isolation and in the context of the whole protein. Comparisons of fold-switching fragments are colored (green, JPRED; blue, PSIPRED; orange, SPIDER2), while comparisons of non-fold-switching fragments from corresponding predictors are gray.

Given that secondary structure predictions depend heavily on homology-based sequence inferences, we hypothesized that removing sequence context from fold-switching protein regions could result in alternative secondary structure predictions. Thus, we ran secondary structure predictions on the amino acid sequences encoding FPRs without flanking NFPR sequences. As expected, predictions of isolated FPR fragments differed from predictions of FPRs in the context of the whole protein. Furthermore, these discrepancies were much larger for FPR fragments than for randomly-selected NFPR fragments of equivalent length (**Fig. 3a-c**), with statistical significances of $1.3*10^{-4}$ (JPRED), $2.2*10^{-2}$ (PSIPRED), and $3.7*10^{-7}$ (SPIDER2). These results indicate that discrepancies in secondary structure predictions of protein fragments in isolations vs. the context of the whole protein are, indeed, a good initial predictor of fold switching.

Discussion

While most proteins with solved structures adhere to the classical notion that proteins adopt one secondary structure scaffold that performs one specific function, there are a number of exceptions [7, 17]. Fold-switching proteins are one class of exceptions because they remodel their secondary structures in response to cellular stimuli, fostering changes in function or enabling tight cellular control. Because fold-switching proteins do not conform to the classical notion, the principles that underlie homology-based secondary structure predictions do not apply to them. Here, we seek to leverage this observation by using faulty secondary structure predictions to identify fold-switching proteins. Similar calculations that collate secondary structure predictions to identify flexible regions in proteins have been performed previously, though on a very limited dataset [18].

Here we show that inaccurate and inconsistent secondary structure predictions can reveal fold switching protein regions. We first showed that secondary structure predictions that are inconsistent with experimentally determined protein structures are significantly more common in FSRs than in non-FSRs. This observation gives statistical power to our previous observation that secondary structure predictions of FSRs are often inconsistent with experiment. We then suggested a new predictive approach by showing that secondary structure predictions of FSRs are often context-dependent. That is, secondary structure predictions of FSRs in isolation often differ from those in context. These context-dependent differences occur significantly more frequently in FSRs than in NFSRs. Thus, these differences can be used as a preliminary indicator of fold switching that relies on amino acid sequence alone, as opposed to previous calculations, which also require one solved structure from a homolog.

In retrospect, it is not surprising that secondary structure predictions of FSRs are context-dependent. A number of computational and experimental studies have highlighted that structural context can determine secondary structure. Pioneering work by Minor and Kim demonstrated that a

"chameleon sequence" of eight amino acids could fold into an α-helix in one part of protein G, but a β-hairpin in another [19]. Since then, several computational studies have scoured the PDB to find dozens of naturally-occurring chameleon sequences, with an upper limit of 20 residues [20-22]. Furthermore, a recent study leveraged context dependence to design 56-amino-acid sequences with 80% sequence identity but different folds [23].

Although context-dependent secondary structure predictions are a good preliminary indicator fold switching, they are not definitive: a small population of NFSRs also show discrepancies between isolated and in-context predictions. Other factors—such as independent folding cooperativity—appear necessary for proteins to switch folds [7]. Furthermore, current secondary structure prediction algorithms depend heavily on available amino acid sequences and proteins with solved structure. Our results suggest that this dependence can lead to prediction bias. Protein structure predictions based on physical principles [24]–instead of homology relationships–could circumvent this bias. Thus, we hope that these findings will encourage the improvement of physically-based secondary structure predictions.

Accurate predictions of fold switching could suggest biological mechanisms underlying observed experimental phenomena. For example, some proteins can change their cellular localizations by switching folds [11, 25]. Others require fold switching to change their functions [9, 26]. Thus, predictions suggesting that a protein switches folds could lead to the generation of hypotheses for how uncharacterized biological processes occur. Furthermore, fold switching proteins are potential drug targets because their conformational equilibria can be disrupted by small molecules. Consistent with this statement, the veterinary medicine halofuginone arrests growth of the malaria parasite through a fold switch of its prolyl tRNA synthetase [27, 28]. Thus, predicting whether a protein switches folds could foster the discovery of new biological processes and treatment for illnesses. We hope that the results presented here will help to lay the groundwork for these advances.

Methods

Secondary structure predictions of whole fold-switching proteins

First, we generated libraries of secondary structure predictions for fold-switching proteins. All full sequences of 192 fold-switching protein structures [7], corresponding to two different conformations of 96 fold-switching proteins (i.e. fold-switch pairs), were downloaded from the PDB [3] and saved as individual FASTA [29] files. Separate secondary structure predictions were run using JPRED4, PSIPRED, and SPIDER2. JPRED predictions were run remotely using a publically downloadable scheduler available on the JPRED4 website [14]. PSIPRED and SPIDER2 calculations were run locally using nr as the database for generating position-specific scoring matrices (PSSMs) from PSI-BLAST [30]. Secondary structure predictions from jnetpred (JPRED), .horiz files (PSIPRED), and .spd3 files (SPIDER2) were converted into FASTA format. Each residue was assigned one of three secondary structures: 'H' for helix, 'E' for extended β-strand, and 'C' for coil. All experimentally determined and predicted secondary structures that were neither helix nor extended were classified as coil (including β-turns), except for chain breaks, which were annotates as '-'. JPRED cannot handle sequences with >800 residues. Thus these sequences were pruned before being submitted to JPRED only; pruning occurred on the N-terminus, C-terminus or both N- and C- termini depending if the FSR was nearer to the C-terminus, N-terminus, or middle of the protein, respectively.

Secondary structure prediction accuracy calculations

Secondary structure prediction accuracies were calculated using the Q_{total} (or Q_3) metric [31], in which predicted secondary structure annotations are compared one-by-one with corresponding secondary structures annotations determined by experiment. Predictions were scored as follows: (in)consistent pairwise predictions were given a score of (0)/1. Prediction scores were then summed and normalized by the length of the sequences compared. Chain breaks were excluded from both scoring and normalization. All sequences composed of ≥10% chain breaks or more were excluded.

Secondary structure prediction accuracy distributions

Distributions of FSRs and NFSRs were calculated as follows. Prediction accuracies were calculated on sequence regions as a sliding window in steps of one. Window size equaled the length of the FSR, as defined by [7], unless that length was <40 residues. FSR lengths < 40 residues were padded symmetrically, if possible, so that the 40-residue threshold was reached. All protein regions containing ≥50% of the fold-switching region was considered an FSR; all others were considered an NFSR. FSRs were determined using the pairwise2.align.localxs function from Biopython [32] with gap-forming score of -1 and gap-elongation score of -0.5. To generate **Fig. 2,** these calculations were run exactly as before, but they excluded the structure from each fold-switch pair with the lower Q_3. All distributions were plotted using Matplotlib [33].

Randomly-generated secondary structure predictions

First, we used the procedure described in *Secondary structure predictions of whole fold-switching proteins* to predict the secondary structures of 226 proteins with high likelihood of not switching folds [7]. Segments of these proteins were randomly selected 10 times each. Segment lengths were randomly selected from a distribution of FSR lengths from the 192 proteins described previously. Secondary structure prediction accuracies were calculated using the Q_3 metric as described previously.

Kolmogorov-Smirnov (KS) statistics and PDB Bias

We found the KS test to give implausibly low p-values for large distributions. To offset this effect, we performed the test using the size of the smaller distribution twice, instead of using the sizes of the smaller and larger distributions once each.

We performed an exhaustive BLAST search of all structures in the PDB and counted structures adopting each conformation from a fold-switch pair by comparing the experimentally-determined secondary structure similarities of their FSRs. PDBs were considered to adopt the same

conformation as the member of the fold switch pair to which they had the highest secondary structure similarity.

Secondary structure distributions of isolated FSRs and NFSRs

Libraries of FSR fragments, padded to 40 residues when necessary, were generated as described previously. Libraries of NFSR fragments were generated by excising 10 random fragments from each non-fold-switching protein with lengths randomly selected from the FSR length distribution described previously. Secondary structure predictions on both libraries were run as described previously. N/FSR distributions were generated by comparing the secondary structure predictions of isolated N/FSR fragments with corresponding secondary structure predictions of N/FSR fragments in context. Only FSRs whose contextualized experimental accuracies were both < 0.8 were included in the distributions.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223-230.
2. Mittermaier A & Kay LE (2006) New tools provide new insights in NMR studies of protein dynamics. *Science* 312(5771):224-228.
3. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242.
4. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85-94.
5. Rost B (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* 266:525-539.
6. Roy A, Kucukural A, & Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725-738.
7. Porter LL & Looger LL (2018) Extant fold-switching proteins are widespread. *Proc Natl Acad Sci U S A* 115(23):5968-5973.
8. Bryan PN & Orban J (2010) Proteins that switch folds. *Curr Opin Struct Biol* 20(4):482-488.
9. Burmann BM, *et al.* (2012) An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* 150(2):291-303.
10. Al Khamici H, *et al.* (2015) Members of the chloride intracellular ion channel protein family demonstrate glutaredoxin-like enzymatic activity. *PLoS One* 10(1):e115699.
11. Littler DR, *et al.* (2004) The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem* 279(10):9298-9305.
12. Ambroggio XI & Kuhlman B (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* 128(4):1154-1161.
13. Porter LL & Rose GD (2012) A thermodynamic definition of protein domains. *Proc Natl Acad Sci U S A* 109(24):9420-9425.

14. Drozdetskiy A, Cole C, Procter J, & Barton GJ (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43(W1):W389-394.
15. McGuffin LJ, Bryson K, & Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404-405.
16. Yang Y*, et al.* (2017) SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. *Methods Mol Biol* 1484:55-63.
17. Wright PE & Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 16(1):18-29.
18. Young M, Kirshenbaum K, Dill KA, & Highsmith S (1999) Predicting conformational switches in proteins. *Protein Sci* 8(9):1752-1764.
19. Minor DL, Jr. & Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature* 380(6576):730-734.
20. Cohen BI, Presnell SR, & Cohen FE (1993) Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci* 2(12):2134-2145.
21. Mezei M (1998) Chameleon sequences in the PDB. *Protein Eng* 11(6):411-414.
22. Mezei M (2018) Revisiting Chameleon Sequences in the Protein Data Bank. *Algorithms* 11(8):N. PAG.
23. Porter LL, He Y, Chen Y, Orban J, & Bryan PN (2015) Subdomain interactions foster the design of two protein pairs with approximately 80% sequence identity but different folds. *Biophys J* 108(1):154-162.
24. Srinivasan R & Rose GD (1999) A physical basis for protein secondary structure. *Proc Natl Acad Sci U S A* 96(25):14258-14263.
25. Xu M*, et al.* (2005) Disulfide isomerization after membrane release of its SAR domain activates P1 lysozyme. *Science* 307(5706):113-117.
26. Chang YG*, et al.* (2015) Circadian rhythms. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science* 349(6245):324-328.
27. Jain V, Kikuchi H, Oshima Y, Sharma A, & Yogavel M (2014) Structural and functional analysis of the anti-malarial drug target prolyl-tRNA synthetase. *J Struct Funct Genomics* 15(4):181-190.
28. Jain V*, et al.* (2015) Structure of Prolyl-tRNA Synthetase-Halofuginone Complex Provides Basis for Development of Drugs against Malaria and Toxoplasmosis. *Structure* 23(5):819-829.
29. Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183:63-98.
30. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
31. Rost B & Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232(2):584-599.
32. Cock PJ*, et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422-1423.
33. Hunter JD (2007) Matplotlib: A 2D graphics environment. *Comput Sci Eng* 9(3):90-95.