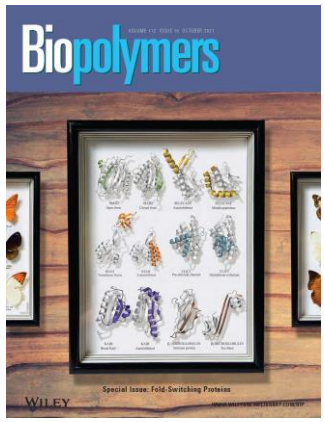
Biopolymers

Special Issue: Fold-Switching Proteins

Guest Editors: Andy LiWang, Lauren L. Porter and Lee-Ping Wang



Received: 5 January 2021 | Revised: 10 August 2021 | Accepted: 10 August 2021

DOI: 10.1002/bip.23471

ARTICLE

Biopolymers WILEY

A sequence-based method for predicting extant fold switchers that undergo α-helix ↔ β-strand transitions

Soumya Mishra[1,3] | Loren L. Looger[3] | Lauren L. Porter[1,2]

[1]National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

[2]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA

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

Correspondence
Lauren L. Porter, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
Email: lauren.porter@nih.gov

Funding information
Howard Hughes Medical Institute; U.S. National Library of Medicine; National Institutes of Health

Abstract

Extant fold-switching proteins remodel their secondary structures and change their functions in response to cellular stimuli, regulating biological processes and affecting human health. Despite their biological importance, these proteins remain understudied. Predictive methods are needed to expedite the process of discovering and characterizing more of these shapeshifting proteins. Most previous approaches require a solved structure or all-atom simulations, greatly constraining their use. Here, we propose a high-throughput sequence-based method for predicting extant fold switchers that transition from α-helix in one conformation to β-strand in the other. This method leverages two previous observations: (a) α-helix ↔ β-strand prediction discrepancies from JPred4 are a robust predictor of fold switching, and (b) the fold-switching regions (FSRs) of some extant fold switchers have different secondary structure propensities when expressed by themselves (isolated FSRs) than when expressed within the context of their parent protein (contextualized FSRs). Combining these two observations, we ran JPred4 on 99-fold-switching proteins and found strong correspondence between predicted and experimentally observed α-helix ↔ β-strand discrepancies. To test the overall robustness of this finding, we randomly selected regions of proteins not expected to switch folds (single-fold proteins) and found significantly fewer predicted α-helix ↔ β-strand discrepancies. Combining these discrepancies with the overall percentage of predicted secondary structure, we developed a classifier to identify extant fold switchers (Matthews correlation coefficient of .71). Although this classifier had a high false-negative rate (7/17), its false-positive rate was very low (2/136), suggesting that it can be used to predict a subset of extant fold switchers from a multitude of available genomic sequences.

KEYWORDS
fold-switching proteins, metamorphic proteins, protein folding, bioinformatics

1 | INTRODUCTION

Extant fold-switching proteins remodel their secondary structures and change their functions in response to cellular stimuli.[1] These environmentally responsive shapeshifters perform over 30 diverse functions, occur in all domains of life, and are associated with diseases such as cancer,[2] autoimmune disorders,[3] and malaria.[4] Furthermore, increasing evidence suggests that extant fold switchers regulate biological processes[5] such as cyanobacterial circadian rhythms[6] and transcription/translation of bacterial virulence genes.[7]

Compared with single-fold proteins, which maintain stable secondary and tertiary structures and typically perform one biological function, extant fold switchers are understudied. Specifically, out of the ~160 000 proteins with solved structures available in the Protein Data Bank (PDB[8]), fewer than 100 have been shown to switch folds. Increasing evidence suggests that fold switching is likely more widespread than currently appreciated,[1] but the current shortage of experimental examples makes it difficult to determine either the physical-chemical properties or the functional scope of fold switchers. Thus, predictive tools are needed to identify more.

Recent computational studies suggest that fold switching is predictable, a prospect that—if realized—could greatly expand the small pool of experimentally determined fold switchers currently available. For example, naturally occurring extant fold switchers were predicted blindly by searching for differences between predicted and experimentally determined protein structures.[1,9] Furthermore, several fold-switching proteins have been designed computationally using the Rosetta software suite.[10,11] Progress has also been made in predicting mutation-induced fold switching[12,13] as well as other conformational changes, such as rigid body motions.[14] Finally, a classifier for extant fold switchers was recently developed as a proof of concept that fold switching is predictable from protein sequence.[15] This classifier is based on confidences of all secondary structure predictions (helix, strand, and coil), whereas the one we developed relies on discrepancies between predicted α-helices and β-strands.

Here we present a sequence-based method for predicting extant fold switchers. This method builds on our previous approach designed for evolved fold switchers, which are defined to have highly similar sequences but different folds.[12] By contrast, extant fold switchers have one sequence that can assume more than one stable secondary and tertiary structure configuration. Whereas the approach for extant fold switchers compared secondary structure predictions of two (or more) different proteins with slightly different sequences, the current method identifies extant fold switchers from the secondary structure predictions of different regions from a single amino acid sequence. The following hypothesis provides the basis for our method: the JPred4 secondary structure prediction of an isolated fold-switching region (FSR) sequence might differ from the JPred4 prediction of the same FSR within the context of its naturally occurring sequence (hereafter called a contextualized FSR). We developed this hypothesis using the previous observation[1] that extant fold-switching proteins generally have: (a) regions that change secondary structure between the two forms (FSRs) and (b) regions that maintain the same secondary structure (structurally constant regions, or SCRs[16]). By definition, FSRs assume multiple stable secondary structures, and several studies have suggested that at least one FSR conformation is stabilized by exogenous interactions.[17,18] Together, these observations indicate that the dominant secondary structure of a given FSR might differ depending on the context of its sequence. Thus, we tested our approach on 99 extant fold switchers with the aim of developing a classifier that could distinguish extant fold switchers from single-folding proteins.

2 | METHODS

2.1 | Selection of extant fold switchers

We selected 93 extant fold switchers from a previous dataset.[1] We excluded 2GED/1NRJB and 3VO9B/3VPAA because they had nearly identical structures but were misclassified due to missing crystal density. We also excluded 1MBYA/2N19A because they come from different organisms, their FSRs differ by three amino acids, and their resting states appear to assume different conformations. Thus, they appear to be evolved—rather than extant—fold switchers. In addition to these 93 extant fold switchers, we included the bacterial cell-division protein MinE,[19,20] SARS-CoV-2 ORF9b,[21] and the human apoptosis regulator BAX,[22] which have all been shown to switch folds, as well as 3 KaiB homologs presumed to switch folds since they come from cyanobacterial strains similar to *Synechococcus elongatus*.

2.2 | JPred4 predictions of extant fold switchers

All amino acid sequences from 99 extant fold switchers with solved structures were downloaded from the PDB and saved as individual FASTA files. JPred4 predictions were run remotely using a publicly downloadable scheduler available on the JPred4 website (http://compbio.dundee.ac.uk/jpred/), and jnetpred predictions were used for all calculations. Jnetpred maximizes accuracy by combining sequence profiles from HMMer[23] and PSI-BLAST,[24] and we found previously that it identifies fold switchers more robustly than other secondary structure predictors.[12] Each residue was assigned one of three secondary structures: "H" for helix, "E" for extended β-strand, and "C" for coil. Chain breaks were annotated "—". PDB IDs and chains of each fold-switched pair, as well as their FSR boundaries, are reported in Table S1. FSR boundaries were initially chosen based on the regions reported previously (bold sequences in table S2 of Ref. [1]). PimA, KaiB, and RfaH were shortened to yield secondary structure prediction discrepancies, and an additional 11 residues were also added to the N-terminal end of PimA's FSR. Such modifications seemed reasonable since JPred4 makes predictions based on a 20-residue window[25] that it could use to associate an isolated fragment with its contextualized secondary structure prediction. Thus, modifying short stretches of N- and C-terminal sequence could decrease the association between isolated sequences and their contextualized predictions.

2.3 | Observed secondary structure discrepancies

Secondary structure classifications of the 93 extant fold switchers were taken from Ref.[1], and classifications of the three KaiB variants were presumed to be the same as *S. elongatus* KaiB. Classifications of MinE, ORF9b, and BAX were determined using DSSP.[26] To quantify secondary structure difference, FSR sequences were aligned with their parents using Biopython[27] pairwise2.align.localxs with gap open/extension penalties of −1.0/−0.5. Secondary structure classifications in the same

register as the aligned FSR sequences were extracted from both experimentally determined structures. Helix ↔ strand discrepancies between the classifications were summed residue-by-residue (1 for discrepancy, 0 for no discrepancy) and normalized by FSR length. Pearson correlations were calculated using the corcoef function from Numpy,[28] and linear fits were determined using Scipy[29] stats.linregress. Our benchmark set was selected by maximizing:

$$\frac{TP^2}{Total},$$

where TP is the number of true positives and "total" is the total number of proteins (true positives + false negatives). Since all 99 proteins switch folds, correct predictions were true positives and incorrect ones were false negatives.

2.4 | Single-fold proteins and fragments

Proteins expected not to switch folds and having fewer than 800 residues (the upper limit in JPred4), totaling 211, were taken from table S3C of Ref. [1]. One segment was selected from a random region of each protein. Segment lengths were randomly selected from a distribution of FSR lengths ranging from 20 to 41, the range of lengths in our benchmark set. Random selections were performed using the *random* module of Python 2.7. JPred4 was run on all 422 sequences (211 full sequences + 211 segments) using its mass-submit scheduler (http://www.compbio.dundee.ac.uk/jpred4/api.shtml#massSubmit).

2.5 | Helix ↔ strand discrepancies and distribution

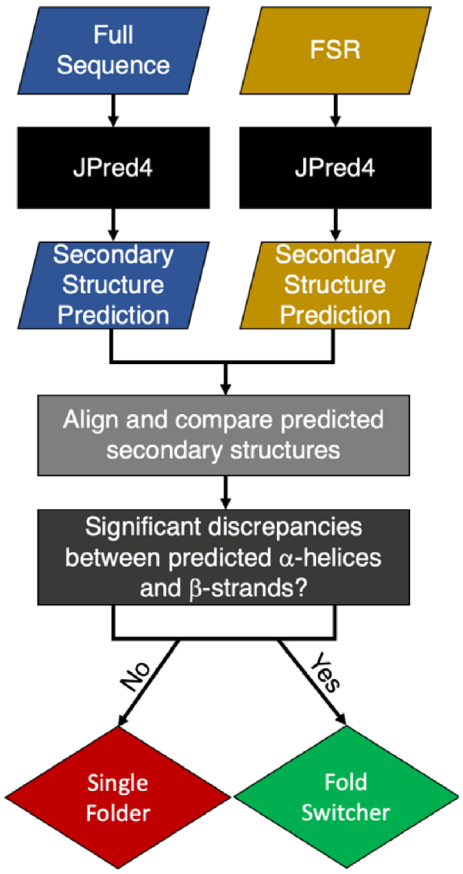
Sequences of isolated FSRs were aligned with full-length proteins using the pairwise2.align.localxs function from Biopython[27] with gap open/extension penalties of −1.0/−0.5. Secondary structure predictions were re-registered according to the resulting alignments and compared. Helix ↔ strand discrepancies between the predictions were summed residue-by-residue (1 for discrepancy, 0 for no discrepancy) and normalized by FSR length. An overall view of our predictive method (Sections 2.2-2.5) is shown in Scheme 1.

2.6 | Distributions and statistics

The distributions in Figures 1 and 3 were generated with Matplotlib.[30] Matthews correlation coefficients[31] were calculated as follows:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP = number of true positives, TN = number of true negatives, FP = number of false positives, and FN = number of false negatives.



SCHEME 1 Summary of predictive approach

2.7 | Chameleon sequences

All 8-residue chameleon sequences (stringent criterion) with non-homologous sequences from the ChSeq[32] database were tested for fold switching. Since JPred4 cannot predict the secondary structures of sequences so short, we extracted 30-residue (mean FSR length of 28 rounded to the nearest multiple of 5) fragments from their parents centered on the chameleon sequences (or as close as possible if the sequences were near termini). JPred4 was then run on all fragments and whole sequences using the mass-submit scheduler. Predictions of whole sequences and fragments were compared as in Section 2.5.

3 | RESULTS

3.1 | JPred4 predicts fold switchers that undergo α-helix ↔ β-strand transitions

We sought to determine whether JPred4 can identify FSRs of extant fold switchers. To do this, JPred4 predictions of isolated FSR sequences and FSRs within their parent sequences (hereafter called contextualized FSRs, Methods) were compared for 99 experimentally validated fold switchers. A moderate Pearson correlation (.67) was

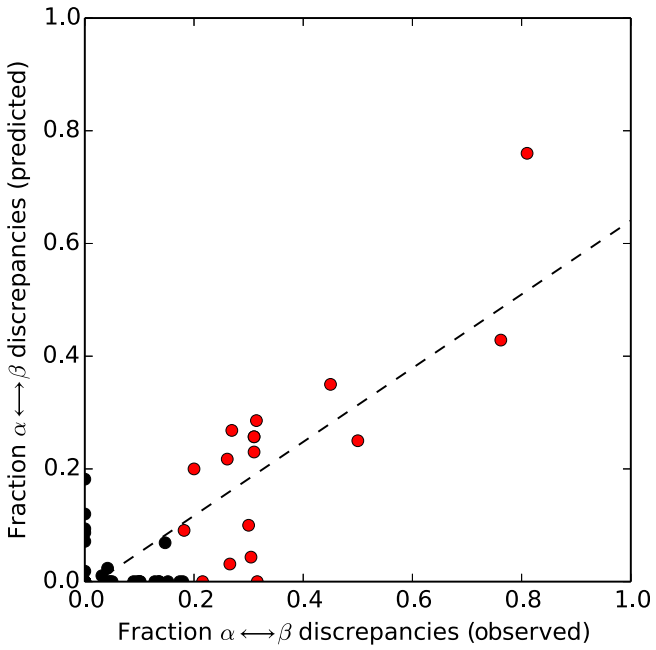


FIGURE 1 Helix ↔ strand discrepancies predicted by JPred4 correspond to experimentally observed α-helix ↔ β-strand differences in fold-switching regions. Dotted line represents best linear fit of all datapoints (black and red circles; Pearson correlation: .82). Red circles correspond to benchmark set of 17-fold switchers. Only 16 can be observed because two KaiB variants (4KSO and 1WWJ) overlap exactly at (0.31, 0.26)

observed between predicted and experimentally observed α-helix ↔ β-strand discrepancies (Figure S1), indicating that JPred4 can identify some fold switchers that undergo α-helix ↔ β-strand transitions. False positives with no observed α ↔ β transitions were eliminated by removing fragments with high levels of predicted coil (≥65%), improving the overall correlation substantially (.82, Figure 1). Together, these results indicate that our method can effectively identify some fold switchers that undergo α-helix ↔ β-strand transitions, but not fold switchers that undergo other types of secondary structure transitions, such as shifts in β-sheet register.

3.2 | Extant fold switchers with sizeable α-helix ↔ β-sheet transitions

We selected a benchmark set of 17 fold switchers by determining the fraction of observed α ↔ β discrepancies that maximized both the percentage and the total number of true positives (Methods, fraction = 0.17, Figure 1, Figure S2). Ten members of this set are highlighted briefly in Figure 2, and all are reported in Table S2.

- Selecase ("selective and specific caseinolytic metallopeptidase"; Figure 2A), produced by archaea and bacteria, and most studied from the archaeon *Methanocaldococcus jannaschii*, is an active metallopeptidase in its monomeric form. Upon forming structured higher-order oligomers, namely dimers, tetramers, and octamers, Selecase is inactivated.[35] Its structures and activities are regulated by its concentration: mostly monomers at 0-0.3 mg/mL; dimers at 0.3-2 mg/mL; tetramers at 2-6 mg/mL, and octamers at >6 mg/mL.
- RfaH (Figure 2B) regulates the expression of virulence proteins from enterobacteria such as *Escherichia coli*.[36] It has two domains: an N-terminal NGN-binding domain (NTD) and a C-terminal domain (CTD) that switches folds. RfaH's CTD folds into an α-helical bundle that forms a binding interface with the NTD, masking its RNA polymerase (RNAP) binding site. Upon binding both RNAP and a specific DNA consensus sequence, called *ops*, the CTD dissociates from the NTD, unmasking the NTD's RNAP binding site. This binding event also triggers the CTD to reversibly refold into a β-barrel able to bind the integral S10 unit of the ribosome and foster efficient translation.[37] When expressed in isolation, RfaH's CTD folds into a β-barrel with no trace of α-helical content (green structure).[37]
- PimA (Figure 2C) is a membrane-associated bacterial glycosyltransferase (phosphatidyl-*myo*-inositol mannosyltransferase) that initiates the biosynthesis of virulence factors produced by *Mycobacterium tuberculosis*. This enzyme has both a closed GDP-bound form and an open form with reshuffled secondary structure. PimA's FSR is highly conserved in mycobacterial orthologs, and both crystallographic and near-UV CD evidence indicate that its open form could play an important role in membrane interactions.[38]
- KaiB (Figure 2D) is a major component of the cyanobacterial circadian clock of *S. elongatus*.[6] Unlike most other circadian clocks, which are driven by transcription-translation oscillation, the cyanobacterial circadian clock is maintained through a periodic phosphorylation cycle, known as a post-translational oscillator (PTO).[39] At night, KaiB's active monomeric form helps to regulate the dephosphorylation of the PTO, while in the morning it primarily populates an inactive tetramer with a different fold, allowing phosphorylation of the PTO.
- Ovalbumin (Figure 2E) is a member of the serpin family (serine protease inhibitor; although ovalbumin is not known to have in situ inhibitory activity—it constitutes 60%-65% of egg whites and appears to be a storage protein[40]) with a zymogenic form (i.e., an inactive precursor, as has plasmepsin). Specifically, inactive ovalbumin has a reactive center loop (RCL) that, when cleaved by a serine protease such as subtilisin, forms a β-strand inserted between two pairs of β-hairpins on its surface. Additionally, the α-helix formed by ovalbumin's uncleaved RCL is regular and less flexible than the distorted helices of inhibitory serpins.[41]
- MinE (Figure 2F) is part of a three-component protein oscillator that helps to regulate bacterial cell division.[19] In its resting state, MinE forms a homodimer with six β-strands (three from each monomer) and four α-helices (two from each monomer). When bound to MinD, another component of the oscillator, MinE's two central β-strands are extruded from its dimer interface and refold into helices that bind MinD,[20] stimulating MinD's ATPase activity and leading to membrane release.
- ORF9b (Figure 2G) is from the genome of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Expressed in

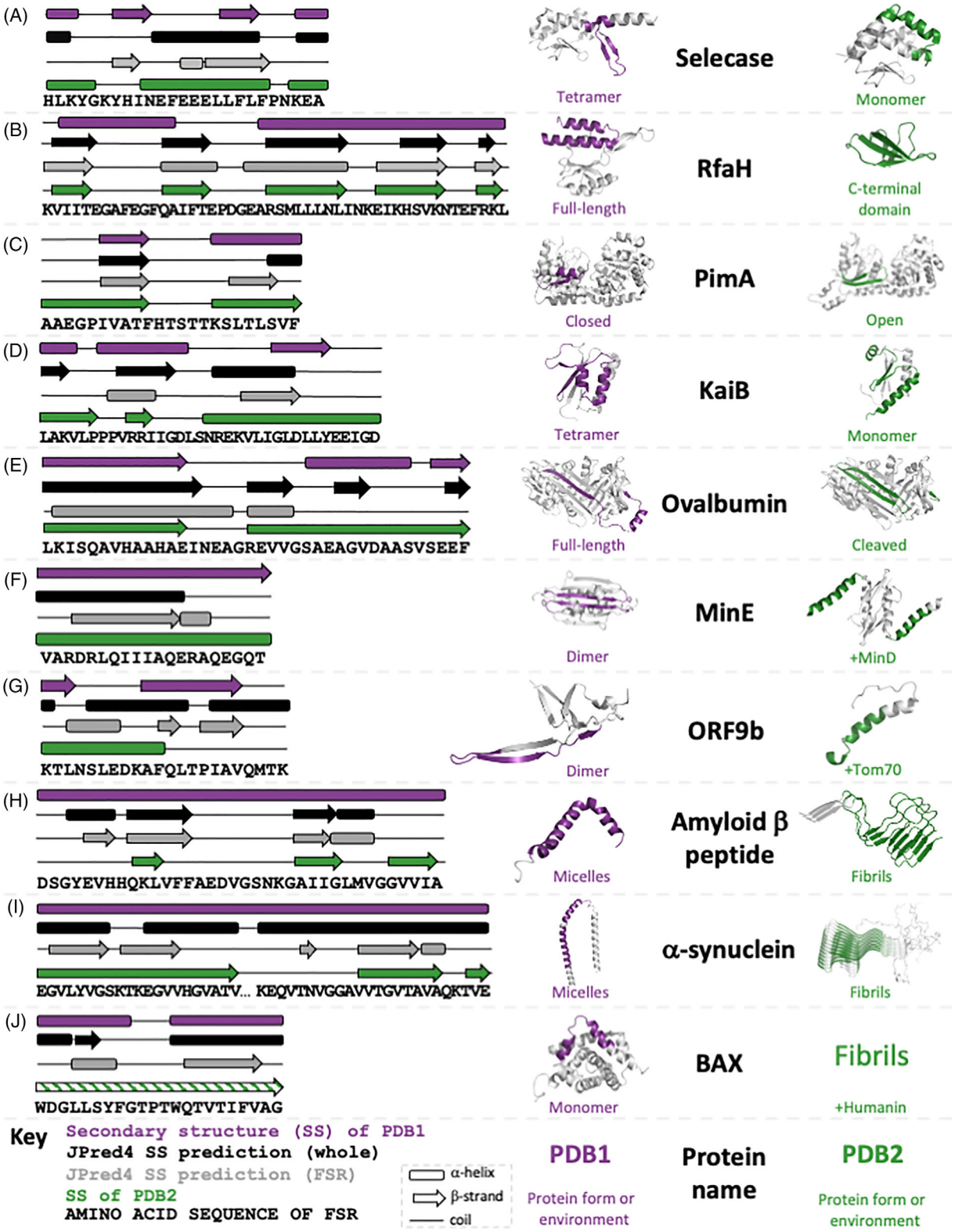


FIGURE 2 JPred4 predicts different secondary structures for isolated and contextualized FSRs of extant fold switchers with substantive transitions between α-helix and β-strand. Each panel shows the experimentally determined secondary structures of both conformations of the fold switcher (purple and green) along with JPred4 secondary structure predictions of the whole sequence (black) and FSR (gray). Purple and green regions of protein structure correspond to FSR sequence shown in diagram; gray corresponds to structurally constant regions (SCRs). Predicted secondary structures that were at least two contiguous residues long are shown. The KaiB variant (2QKE) represents all members of the KaiB family; the other three (1WWJ, 4SKO, 1R5P) are not shown; amylin is also not shown due to lack of space. The differential secondary structure predictions for ORF9b were reported previously.[33] The green secondary structure diagram of BAX is shaded with lines to signify that its structure has not been solved, though other experimental evidence strongly suggests that it folds into a β-sheet. All three-dimensional protein structures were made using PyMOL[34]

isolation, it forms a homodimer composed of β-sheets. When bound to human Tom70, however, it refolds into an α-helix with one of two possible cellular effects[21]: (a) modulating interferon and apoptosis signaling or (b) decreasing mitochondrial import efficiency, leading to mitophagy. JPred4 has been used previously to predict ORF9b's fold switching.[33]

- The human amyloid-forming proteins α-Synuclein and amyloid β (Figure 2H,I, respectively), along with amylin (Table S2), are all believed to interact with membranes, where they form α-helices.[42–44] While the cognate functions of helical α-synuclein and amyloid β remain under investigation, amylin is an endocrine hormone (co-secreted with insulin) that regulates glycemic metabolism.[43] All three peptides can also form fibrillar deposits associated with diseases such as Parkinson's (α-Synuclein),[45] type 2 diabetes (amylin),[46] and Alzheimer's (amyloid β).[47]
- BAX is a human protein involved in mitochondrial apoptosis. It assumes an all α-helical fold in the cytosol, and membrane insertion of its C-terminal helix appears to foster its apoptotic function. Several lines of experimental evidence (e.g., mass spectrometry, electron microscopy, and circular dichroism spectroscopy) indicate that BAX refolds into β-sheet fibrils when bound to the humanin peptide.[22] Furthermore, light-scattering experiments demonstrate that its C-terminal helix propagates fibril formation.[22] This refolding is believed to sequester BAX, preventing it from initiating mitochondrial apoptosis.

In all cases shown in Figure 2, along with 5/7 of the other proteins in our benchmark set (Table S2), we found that JPred4 predicted different secondary structures for isolated and contextualized FSR sequences. JPred4 secondary structure predictions tend to correspond reasonably well (<6% α-helix ↔ β-strand discrepancies[9]) with at least one experimentally determined protein structure for all 14 proteins. In fact, in all 17 cases, α-helix and β-strand secondary structure elements correspond well between one prediction and one experimentally determined conformation (correct secondary structures in all of the right positions, though not necessarily the experimentally determined lengths). However, the alternative JPred4 predictions generally do not correspond well with the alternative secondary structure prediction, except for PimA and KaiB. Nevertheless, as in previous work,[12] we use discrepancies between predictions to infer fold switching; for our purposes, the accuracies of the JPred4 predictions have no bearing on this inference.

3.3 | JPred4 discriminates between FSRs and single folding regions

To determine the significance of JPred4's α-helix ↔ β-strand prediction discrepancies for isolated and contextualized FSRs, we randomly selected fragments from a set of 211 proteins expected not to switch folds (single-fold proteins). Upon eliminating all predictions with ≥65% coil, 136 predictions remained.

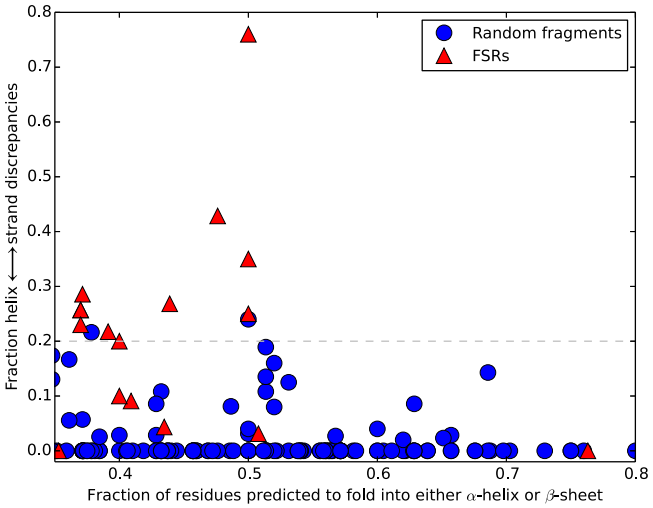


FIGURE 3 JPred4 discriminates between single folders and fold switchers. Single folders/fold switchers are blue circles/red triangles. The dashed line represents the threshold for classifying fold switchers by fraction of predicted secondary structure/fraction of α-helix ↔ β-strand discrepancies (0.2). Datapoints at or above this threshold are predicted to switch folds. Only 16/17-fold switchers can be seen because two KaiB variants have identical coordinates (0.37, 0.26)

Predictions of single folders and fold switchers are compared in Figure 3. We noticed that 11/17 of the fold-switching proteins in our benchmark set had predicted helix ↔ strand discrepancies ≥20%, while only 2/136 of single folders had helix ↔ strand discrepancies at the same threshold. One of these false positives comprised residues 12-48 from the glutathione S-transferase (GST) Omega 3 expressed by the silkworm *Bombyx mori*. Residue 29 of this segment is an asparagine, which replaces a highly conserved cysteine in the other members of the Omega family.[48] This single amino acid change is partially responsible for Omega 3's loss of GST activity: mutating asparagine 29 to a cysteine while also deleting its flexible C-terminal helix restores GST activity. Interestingly, running JPred4 on the same segment (residues 12-48) with just an N29C mutation gives the same secondary structure prediction as that of the whole protein (Table S3).[48] Based on our previous work on sequence-similar fold switchers,[12] this result suggests that this protein segment might switch folds and thus might not be a false positive after all. The other false positive comprised residues 93-142 of Bd3460, a self-protection protein from *Bdellovibrio bacteriovorus* that assumes an ankyrin-like fold.[49] No obvious reason for the fold switching misclassification was identified.

At a 20% threshold for predicted α-helix ↔ β-strand discrepancies, our method yielded 11 true positives, 2 false positives, 134 true negatives, and 6 false negatives, resulting in a Matthews correlation coefficient of .71 (very low false-positive rate; moderate false-negative rate). In 4/6 false negatives, α-helix ↔ β-strand discrepancies were predicted, but they were not large enough to exceed the 20% threshold for the classifier. JPred4 may have misclassified the six false negatives for two reasons. Firstly, we suspect that the sequence

profiles generated for FSRs and whole proteins were similar, leading to identical JPred predictions. Secondly, database population may have played a role in the misclassification. Specifically, sequences associated with 1-fold may have been more highly represented than sequences associated with the other.

3.4 | JPred4 does not systematically classify chameleon sequences as fold switchers

To further test the robustness of our classifier, we ran our JPred4-based method on 45 nonhomologous chameleon sequences from the ChSeq database.[32] Chameleon sequences are identical sequences that assume α-helices in some proteins and β-strands in others but are not associated with fold switching.[50] Of the 36 sequences with <65% coil predicted, 5 were classified as putative fold switchers (Table S4). Thus, while our method sometimes misclassifies chameleon sequences as fold switchers, it is not a systematic defect.

4 | DISCUSSION

Fold switchers are exceptions to the observation that folded proteins assume one stable structure that performs one function. Nevertheless, increasing evidence suggests that these proteins may be more abundant in nature than previously thought.[1] Fold switching impacts protein function[5] and is associated with multiple diseases.[2,3,51] Thus, it would be useful to have a bioinformatic algorithm that identifies more fold switchers from their sequences. This is especially true because, up to this point, all experimentally characterized fold switchers have been stumbled upon by chance.[1]

Here we present an approach for predicting extant fold switchers from their amino acid sequences alone. This method is based on previous experimental work suggesting that the FSRs of proteins are context-dependent: that is, their conformations are determined by their environment.[17,18] In light of this, we hypothesized that it might be possible to predict extant fold switchers by comparing the JPred4 secondary structure predictions of isolated FSRs with contextualized FSRs and searching for α-helix ↔ β-strand discrepancies. Indeed, significant discrepancies were found in 11/17-fold switchers used in this study. We used this finding to develop a classifier for extant fold switchers that yielded a Matthews correlation coefficient of .71. We suspect that JPred4 successfully identified extant fold switchers for the same reason it identified sequence-similar fold switchers[12]: different sequences (contextualized and isolated FSRs in this case) yielded different sequence profiles from PSI-BLAST searches. Future work revealing how these different profiles lead to dramatically different secondary structure predictions would be useful.

Two additional results stand out in light of our previous method,[12] which predicts evolved fold switchers with highly similar sequences. First, the method presented here predicts fold switching in all four KaiB variants tested. This positive result is an improvement over our previous method for sequence-similar fold switchers, which failed to predict fold switching in all KaiB variants.[12] Secondly, our results strongly suggest that the fragment from Omega 3 is an FSR, even though it was in our set of proteins not expected to switch folds. Just one mutation (N29C) is sufficient to dramatically change the secondary structure predictions of this sequence, a previously identified characteristic of sequence-similar fold switchers (proteins with highly similar—but not identical—amino acid sequences and different folds[12]). Additionally, Omega 3's GST topology[48] has been known to switch folds in other proteins, namely KaiB[52] and chloride intracellular channel 1 (CLIC1).[53] Still, further experimental work would be needed to determine whether Omega 3 switches folds.

Although we are optimistic that the approach presented here can be used to predict novel fold switchers, it has several limitations. Firstly, it can only identify fold switchers that undergo large α-helix ↔ β-sheet transitions. To date, these proteins are rare and comprise only 17% of known fold switchers. Biologically important fold switchers like lymphotactin,[54] which maintains β-sheets that change their hydrogen bonding register, and most β-pore proteins,[55] which extend already existing β-sheet structures, will be missed. Secondly, it will not identify all fold switchers that undergo large α-helix ↔ β-sheet transitions, as evidenced by the fact that only 11/17 of the fold switchers tested gave a robust enough signal to be classified positively. Thirdly, because the FSRs of undiscovered fold switchers are not known a priori, our method will likely need to test many putative FSRs (different sizes and different regions) within the same protein to determine whether or not it is a fold switcher. Although this approach is much less computationally intensive than all-atom simulations, it will still require substantial time and computational power to predict fold switching in thousands of genomic sequences. Furthermore, the more sequences probed, the more likely false positives will be hit. Additional work will be needed to more accurately distinguish between these false positives and true fold switchers. Finally, our training set was small, comprising only 17 known fold switchers suitable for the predictive method presented here. Thus, it is likely that our statistics, especially for true positives and false negatives, are noisy. As more fold switchers are discovered, we are optimistic that it will be possible to develop methods that can predict more types of fold switchers with higher accuracy.

5 | CONCLUSIONS

Our results suggest that the α-helix ↔ β-strand transitions of some extant fold switchers can be predicted from their sequences alone using the homology-based secondary structure predictor JPred4. Although this method will not identify all extant fold switchers whose secondary structures transition from α-helix ↔ β-strand, its low false positive (2/136) and moderate true positive (11/17) rates suggest that many positive predictions will likely correspond to true extant fold switchers. Thus, we are optimistic that this approach can be used to predict a subset of extant fold switchers from the broad base of available genomic sequences.

ACKNOWLEDGMENTS

This work utilized the computational resources of the NIH HPS Biowulf cluster (http://hpc.nih.gov). This work was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Github at https://github.com/porterll/extant_fold_switchers

ORCID

Lauren L. Porter https://orcid.org/0000-0003-2031-8326

REFERENCES

[1] L. L. Porter, L. L. Looger, *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 5968.
[2] B. P. Li, Y. T. Mao, Z. Wang, Y. Y. Chen, Y. Wang, C. Y. Zhai, B. Shi, S. Y. Liu, J. L. Liu, J. Q. Chen, *Cell. Physiol. Biochem.* **2018**, *46*, 907.
[3] Y. Lei, Y. Takahama, *Microbes Infect.* **2012**, *14*, 262.
[4] V. Jain, H. Kikuchi, Y. Oshima, A. Sharma, M. Yogavel, *J. Struct. Funct. Genomics* **2014**, *15*, 181.
[5] A. K. Kim, L. L. Porter, *Structure* **2021**, *29*, 6.
[6] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, A. LiWang, *Science* **2015**, *349*, 324.
[7] J. Y. Kang, R. A. Mooney, Y. Nedialkov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, S. A. Darst, *Cell* **2018**, *173*, 1650.
[8] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, C. Zardecki, *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 899.
[9] S. Mishra, L. L. Looger, L. L. Porter, *Protein Sci.* **2019**, *28*, 1487.
[10] X. I. Ambroggio, B. Kuhlman, *J. Am. Chem. Soc.* **2006**, *128*, 1154.
[11] K. Y. Wei, D. Moschidi, M. J. Bick, S. Nerli, A. C. McShan, L. P. Carter, P. S. Huang, D. A. Fletcher, N. G. Sgourakis, S. E. Boyken, D. Baker, *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 7208.
[12] A. K. Kim, L. L. Looger, L. L. Porter, *Biopolymers* **2021**, e23416.
[13] P. Tian, R. B. Best, *PLoS Comput. Biol.* **2020**, *16*, e1008285.
[14] P. Sfriso, M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, M. Orozco, *Structure* **2016**, *24*, 116.
[15] N. Chen, M. Das, A. LiWang, L. P. Wang, *Biophys. J.* **2020**, *119*, 1380.
[16] I. K. Huang, J. Pei, N. V. Grishin, *Bioinformatics* **2013**, *29*, 175.
[17] D. L. Minor Jr., P. S. Kim, *Nature* **1996**, *380*, 730.
[18] L. L. Porter, Y. He, Y. Chen, J. Orban, P. N. Bryan, *Biophys. J.* **2015**, *108*, 154.
[19] M. Cai, Y. Huang, Y. Shen, M. Li, M. Mizuuchi, R. Ghirlando, K. Mizuuchi, G. M. Clore, *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 25446.
[20] K. T. Park, W. Wu, K. P. Battaile, S. Lovell, T. Holyoak, J. Lutkenhaus, *Cell* **2011**, *146*, 396.
[21] D. E. Gordon, J. Hiatt, M. Bouhaddou, V. V. Rezelj, S. Ulferts, H. Braberg, A. S. Jureka, K. Obernier, J. Z. Guo, J. Batra, R. M. Kaake, A. R. Weckstein, T. W. Owens, M. Gupta, S. Pourmal, E. W. Titus, M. Cakir, M. Soucheray, M. McGregor, Z. Cakir, G. Jang, M. J. O'Meara, T. A. Tummino, Z. Zhang, H. Foussard, A. Rojc, Y. Zhou, D. Kuchenov, R. Huttenhain, J. Xu, M. Eckhardt, D. L. Swaney, J. M. Fabius, M. Ummadi, B. Tutuncuoglu, U. Rathore, M. Modak, P. Haas, K. M. Haas, Z. Z. C. Naing, E. H. Pulido, Y. Shi, I. Barrio-Hernandez, D. Memon, E. Petsalaki, A. Dunham, M. C. Marrero, D. Burke, C. Koh, T. Vallet, J. A. Silvas, C. M. Azumaya, C. Billesbolle, A. F. Brilot, M. G. Campbell, A. Diallo, M. S. Dickinson, D. Diwanji, N. Herrera, N. Hoppe, H. T. Kratochvil, Y. Liu, G. E. Merz, M. Moritz, H. C. Nguyen, C. Nowotny, C. Puchades, A. N. Rizo, U. Schulze-Gahmen, A. M. Smith, M. Sun, I. D. Young, J. Zhao, D. Asarnow, J. Biel, A. Bowen, J. R. Braxton, J. Chen, C. M. Chio, U. S. Chio, I. Deshpande, L. Doan, B. Faust, S. Flores, M. Jin, K. Kim, V. L. Lam, F. Li, J. Li, Y. L. Li, Y. Li, X. Liu, M. Lo, K. E. Lopez, A. A. Melo, F. R. Moss 3rd., P. Nguyen, J. Paulino, K. I. Pawar, J. K. Peters, T. H. Pospiech Jr., M. Safari, S. Sangwan, K. Schaefer, P. V. Thomas, A. C. Thwin, R. Trenker, E. Tse, T. K. M. Tsui, F. Wang, N. Whitis, Z. Yu, K. Zhang, Y. Zhang, F. Zhou, D. Saltzberg, Q. S. B. Consortium, A. J. Hodder, A. S. Shun-Shion, D. M. Williams, K. M. White, R. Rosales, T. Kehrer, L. Miorin, E. Moreno, A. H. Patel, S. Rihn, M. M. Khalid, A. Vallejo-Gracia, P. Fozouni, C. R. Simoneau, T. L. Roth, D. Wu, M. A. Karim, M. Ghoussaini, I. Dunham, F. Berardi, S. Weigang, M. Chazal, J. Park, J. Logue, M. McGrath, S. Weston, R. Haupt, C. J. Hastie, M. Elliott, F. Brown, K. A. Burness, E. Reid, M. Dorward, C. Johnson, S. G. Wilkinson, A. Geyer, D. M. Giesel, C. Baillie, S. Raggett, H. Leech, R. Toth, N. Goodman, K. C. Keough, A. L. Lind, C. Zoonomia, R. J. Klesh, K. R. Hemphill, J. Carlson-Stevermer, J. Oki, K. Holden, T. Maures, K. S. Pollard, A. Sali, D. A. Agard, Y. Cheng, J. S. Fraser, A. Frost, N. Jura, T. Kortemme, A. Manglik, D. R. Southworth, R. M. Stroud, D. R. Alessi, P. Davies, M. B. Frieman, T. Ideker, C. Abate, N. Jouvenet, G. Kochs, B. Shoichet, M. Ott, M. Palmarini, K. M. Shokat, A. Garcia-Sastre, J. A. Rassen, R. Grosse, O. S. Rosenberg, K. A. Verba, C. F. Basler, M. Vignuzzi, A. A. Peden, P. Beltrao, N. J. Krogan, *Science* **2020**, *370*, eabe9403.
[22] D. L. Morris, D. W. Kastner, S. Johnson, M. P. Strub, Y. He, C. K. E. Bleck, D. Y. Lee, N. Tjandra, *J. Biol. Chem.* **2019**, *294*, 19055.
[23] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, R. D. Finn, *Nucleic Acids Res.* **2018**, *46*, W200.
[24] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
[25] A. Drozdetskiy, C. Cole, J. Procter, G. J. Barton, *Nucleic Acids Res.* **2015**, *43*, W389.
[26] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.
[27] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. de Hoon, *Bioinformatics* **2009**, *25*, 1422.
[28] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Rio, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, *Nature* **2020**, *585*, 357.
[29] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, C. SciPy, *Nat. Methods* **2020**, *17*, 261.
[30] J. D. Hunter, *Comput. Sci. Eng.* **2007**, *9*, 90.
[31] B. W. Matthews, *Biochim. Biophys. Acta* **1975**, *405*, 442.
[32] W. Li, L. N. Kinch, P. A. Karplus, N. V. Grishin, *Protein Sci.* **2015**, *24*, 1075.
[33] L. L. Porter, *Protein Sci.* **2021**, *30*, 1723.
[34] The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
[35] M. Lopez-Pelegrin, N. Cerda-Costa, A. Cintas-Pedrola, F. Herranz-Trillo, P. Bernado, J. R. Peinado, J. L. Arolas, F. X. Gomis-Ruth, *Angew. Chem. Int. Ed. Engl.* **2014**, *53*, 10624.
[36] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, P. Rosch, *Cell* **2012**, *150*, 291.
[37] P. K. Zuber, K. Schweimer, P. Rosch, I. Artsimovitch, S. H. Knauer, *Nat. Commun.* **2019**, *10*, 702.

[38] D. Giganti, D. Albesa-Jove, S. Urresti, A. Rodrigo-Unzueta, M. A. Martinez, N. Comino, N. Barilone, M. Bellinzoni, A. Chenal, M. E. Guerin, P. M. Alzari, *Nat. Chem. Biol.* **2015**, *11*, 16.
[39] C. L. Partch, *J. Mol. Biol.* **2020**, *432*, 3426.
[40] P. E. Stein, A. G. Leslie, J. T. Finch, R. W. Carrell, *J. Mol. Biol.* **1991**, *221*, 941.
[41] M. Yamasaki, Y. Arii, B. Mikami, M. Hirose, *J. Mol. Biol.* **2002**, *315*, 113.
[42] O. Crescenzi, S. Tomaselli, R. Guerrini, S. Salvadori, A. M. D'Ursi, P. A. Temussi, D. Picone, *Eur. J. Biochem.* **2002**, *269*, 5642.
[43] S. M. Patil, S. Xu, S. R. Sheftic, A. T. Alexandrescu, *J. Biol. Chem.* **2009**, *284*, 11982.
[44] J. N. Rao, C. C. Jao, B. G. Hegde, R. Langen, T. S. Ulmer, *J. Am. Chem. Soc.* **2010**, *132*, 8657.
[45] M. A. Walti, F. Ravotti, H. Arai, C. G. Glabe, J. S. Wall, A. Bockmann, P. Guntert, B. H. Meier, R. Riek, *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E4976.
[46] Q. Cao, D. R. Boyer, M. R. Sawaya, P. Ge, D. S. Eisenberg, *Nat. Struct. Mol. Biol.* **2020**, *27*, 653.
[47] M. D. Tuttle, G. Comellas, A. J. Nieuwkoop, D. J. Covell, D. A. Berthold, K. D. Kloepper, J. M. Courtney, J. K. Kim, A. M. Barclay, A. Kendall, W. Wan, G. Stubbs, C. D. Schwieters, V. M. Lee, J. M. George, C. M. Rienstra, *Nat. Struct. Mol. Biol.* **2016**, *23*, 409.
[48] B. Y. Chen, X. X. Ma, P. C. Guo, X. Tan, W. F. Li, J. P. Yang, N. N. Zhang, Y. Chen, Q. Xia, C. Z. Zhou, *J. Mol. Biol.* **2011**, *412*, 204.
[49] C. Lambert, I. T. Cadby, R. Till, N. K. Bui, T. R. Lerner, W. S. Hughes, D. J. Lee, L. J. Alderwick, W. Vollmer, R. E. Sockett, A. L. Lovering, *Nat. Commun.* **2015**, *6*, 8884.
[50] M. Mezei, *Proteins* **2021**, *89*, 3.
[51] V. Jain, M. Yogavel, Y. Oshima, H. Kikuchi, B. Touquet, M. A. Hakimi, A. Sharma, *Structure* **2015**, *23*, 819.
[52] R. Tseng, N. F. Goularte, A. Chavan, J. Luu, S. E. Cohen, Y. G. Chang, J. Heisler, S. Li, A. K. Michael, S. Tripathi, S. S. Golden, A. LiWang, C. L. Partch, *Science* **2017**, *355*, 1174.
[53] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, M. Mazzanti, S. N. Breit, P. M. Curmi, *J. Biol. Chem.* **2004**, *279*, 9298.
[54] A. F. Dishman, R. C. Tyler, J. C. Fox, A. B. Kleist, K. E. Prehoda, M. M. Babu, F. C. Peterson, B. F. Volkman, *Science* **2021**, *371*, 86.
[55] M. Podobnik, P. Savory, N. Rojko, M. Kisovec, N. Wood, R. Hambley, J. Pugh, E. J. Wallace, L. McNeill, M. Bruce, I. Liko, T. M. Allison, S. Mehmood, N. Yilmaz, T. Kobayashi, R. J. Gilbert, C. V. Robinson, L. Jayasinghe, G. Anderluh, *Nat. Commun.* **2016**, *7*, 11598.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: S. Mishra, L. L. Looger, L. L. Porter, *Biopolymers* **2021**, *112*(10), e23471. https://doi.org/10.1002/bip.23471