**A sequence-based method for predicting extant fold switchers
that undergo α -helix $\leftrightarrow$ β -strand transitions**

Soumya Mishra^{1,3}, Loren L. Looger³, and Lauren L. Porter^{1,2,3,*}

¹National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

²National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892,
USA

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

*Correspondence: lauren.porter@nih.gov

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. In spite of their biological importance, these proteins remain understudied. Few representative examples of fold switchers are available in the Protein Data Bank, and they are difficult to predict. In fact, all 96 experimentally validated examples of extant fold switchers were stumbled upon by chance. Thus, predictive methods are needed to expedite the process of discovering and characterizing more of these shapeshifting proteins. 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: (1) α -helix $\leftrightarrow$ β -strand prediction discrepancies from JPred4 are a robust predictor of fold switching, and (2) the fold-switching regions (FSRs) of some extant fold switchers have different secondary structure propensities when expressed in isolation (isolated FSRs) than when expressed within the context of their parent protein (contextualized FSRs). Combining these two observations, we ran JPred4 on the sequences of isolated and contextualized FSRs from 14 known extant fold switchers and found α -helix $\leftrightarrow$ β -strand prediction discrepancies in every case. 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 α -helix $\leftrightarrow$ β -strand prediction discrepancies ($p < 4.2 \times 10^{-20}$, Kolmogorov-Smirnov test). Combining these discrepancies with the overall percentage of predicted secondary structure, we developed a classifier that often robustly identifies extant fold switchers (Matthews Correlation Coefficient of 0.70). Although this classifier had a high false negative rate (6/14), its false positive rate was very low (1/211), suggesting that it can be used to predict a subset of extant fold switchers from billions of available genomic sequences.

1. Introduction

Extant fold-switching proteins remodel their secondary structures and change their functions in response to cellular stimuli¹. Known instances of these environmentally responsive shapeshifters perform over 30 diverse functions, occur in all domains of life, and are associated with diseases such as cancer², autoimmune disorders³, and malaria⁴. Furthermore, increasing evidence suggests that extant fold switchers are central to the regulation of biological processes⁵ such as cyanobacterial circadian rhythms⁶ and transcription/translation of bacterial virulence genes⁷.

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 (PDB⁸), fewer than 100 have been shown to switch folds. Increasing evidence suggests that fold switching is likely more widespread than currently appreciated¹, 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, a couple of fold-switching proteins have been designed computationally using the Rosetta software suite^{10, 11}. Finally, progress has been made in predicting mutation-induced fold switching^{12, 13} as well as other conformational changes, such as rigid body motions¹⁴.

To date, all computational methods used to predict and design extant fold switchers suffer from the same drawback: the need for a protein structure on which to base predictions. This drawback severely limits the number of proteins that can be assessed for fold switching. Specifically, there are nearly 190,000,000 non-redundant, well-annotated protein sequences available in the RefSeq database¹⁵ but only 160,000 protein structures available in the PDB—which group into <15,000 unique structures¹⁶. Thus, any method that requires a solved protein structure to make predictions can sample < 0.01-0.1% of the protein sequences available. Furthermore, methods like Rosetta perform atomistic simulations, which are computationally intensive and not suitable for predicting the structures of thousands of proteins in parallel. For these reasons, sequence-based methods are needed to effectively identify fold switchers from the wealth of available genomic sequences.

Here we present a sequence-based method for predicting extant fold switchers that is based on the following hypothesis: 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). The basis for this hypothesis follows. We started from the previous observation¹ that extant fold-switching proteins generally have: (1) regions that change secondary structure between the two forms (FSRs) and (2) regions that maintain the same secondary structure (structurally constant regions, or SCRs¹⁷). By definition, FSRs assume multiple stable secondary structures, and several studies have suggested that at least one FSR conformation is stabilized by exogenous interactions^{18, 19}. Together, these observations indicate that the dominant secondary structure of a given FSR might differ depending on the context of its sequence. Thus, an FSR might assume a different secondary structure when expressed in isolation than when expressed within a larger protein sequence. Indeed, this is known to be the case for the C-terminal domain of RfaH,

which folds into an α -helical bundle when covalently attached to its N-terminal domain but assumes a β -barrel conformation when expressed as an isolated domain²⁰.

Since previous work strongly suggests that fold switching is most robustly predicted using α -helix $\leftrightarrow$ β -sheet prediction discrepancies from JPred4^{13, 21}, we tested our hypothesis on 14 extant fold switchers with substantial α -helix $\leftrightarrow$ β -sheet transitions. In all 14 cases, JPred4 predicted that isolated and contextualized FSR sequences would have different α -helix and β -strand content. Specifically, overlapping regions were predicted to be α -helix for the isolated/contextualized sequence and β -sheet for the other sequence. To test the robustness of this finding, we randomly selected regions within a set of 211 proteins not expected to switch folds and found the secondary structure predictions of isolated and contextualized sequences to be significantly more consistent than they were for fold-switching proteins ($p < 4.2 \times 10^{-20}$, Kolmogorov-Smirnov test). We used this result to develop a classifier for extant fold switchers, which proved to be robust (Matthews Correlation Coefficient = 0.70). These results suggest that our approach can be used to predict a subset of extant fold switchers from the broad base of available genomic sequences.

2. Methods

2.1. Selection of extant fold switchers

Using a previous dataset¹, we selected all extant fold switchers with FSRs ≥ 20 residues (the minimum length for JPred4 predictions) and an FSR region with at least 4 contiguous residues that assume an α -helix in one conformation and a β -strand in the other. This corresponded to a minimum secondary structure difference score of 0.6. (Secondary structure differences were scored as previously: α -helix $\leftrightarrow$ β -strand: 1; coil $\leftrightarrow$ α -helix or coil $\leftrightarrow$ β -strand: 0.5; no change in secondary structure: 0. Overall score is the total score normalized by FSR length.) Secondary structures of all extant fold switchers were classified using DSSP²².

2.2. JPred4 predictions of extant fold switchers

All amino acid sequences from 14 extant fold switchers with solved structures were downloaded from the Protein Data Bank (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. 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 were as follows: PimA_{open}/PimA_{closed}: 4NC9C/4N9WA; Selecase₄/Selecase₁: 4QHHA/4QHFA; RfaH_{transcription}/RfaH_{translation}: 2OUGA, 2LCLA; KaiB₁/KaiB₄ (4 variants): 5JYTA/2QKEA, 1WWJA, 4KSOA, 1R5PA; PfPRS_{apo}/PfPRS_{liganded}: 4YDQA/4TWAA; Ovalbumin_{cleaved}/Ovalbumin_{uncleaved}: 1JTIB/1OVAA; Proplasmepsin/Plasmepsin: 1MIQA/1QS8B; SUN2/SUN2-KASH1: 4DXTA/4DXRA; A β fibril/A β peptide: 2NAOF/1ITYA; Amylin/Amylin fibril: 2KB8A/6VW2A; α -synuclein fibril/ α -synuclein: 2N0AD/2KKWA. All naturally occurring KaiB variants from cyanobacteria were included because we presumed that they all switch folds, though this has only been shown explicitly for *S. elongatus*. One additional hit (human primase—3L9Q/4RR2) was excluded because very recent NMR work strongly suggests that it is not a fold switcher but rather maintains its helical structure in solution; the alternative β -hairpin conformation is likely a crystallographic artifact²³. FSR boundaries were initially chosen based on the regions reported previously (bold sequences in Table S2 of¹). Fragments of KaiB and PimA were shortened by 13 and 17 residues, respectively, to yield secondary structure prediction discrepancies (sequences in Table S1). 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 window²¹ 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. 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 of¹. One segment was selected from a random region of each protein. Segment lengths were randomly selected from the distribution of 14 FSR lengths (the lengths of the sequences in **Table S1**). Random selections were performed using the *random* module of Python2.7. 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>).

2.4 Helix <-> strand discrepancies and distribution

Sequences of isolated FSRs were aligned with full-length proteins using the `pairwise2.align.localxs` function from Biopython²⁴ 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 the minimum number of secondary structure annotations (helix and sheet) in the pair of sequences compared.

2.5. Distributions and statistics

The distributions in **Figures 2 and 3** was generated with Matplotlib²⁵. Matthews Correlation Coefficients²⁶ 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 is the number of false negatives.

3. Results

3.1. Extant fold switchers with sizeable α -helix $\leftrightarrow$ β -sheet transitions

We searched a previously published set of fold-switching proteins for all instances in which FSRs were at least 20 residues long and had at least four contiguous residues forming an α -helix in one conformation formed a β -strand in the other. This yielded 11 unique proteins (**Figure 1**), each which are highlighted briefly:

- Selecase (“selective and specific caseinolytic metallopeptidase”; **Figure 1A**), 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 inactivated²⁷. 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 1B**) regulates the expression of virulence proteins from enterobacteria such as *Escherichia coli*²⁰. 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 translation²⁸. When expressed in isolation, RfaH’s CTD folds into a β -barrel with no trace of α -helical content (green structure)²⁸.

- PimA (**Figure 1C**) 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 interactions²⁹.
- PfPRS (**Figure 1D**) is a homodimeric prolyl tRNA synthetase from the malaria parasite *Plasmodium falciparum*. When bound to the inhibitor halofuginone, the active sites of the PfPRS monomers fold into β -sheets that hydrogen bond to active-site β -sheets of the other monomeric unit³⁰. By contrast, the active site of apo PfPRS forms helical structures⁴.
- KaiB (**Figure 1E**) is a major component of the cyanobacterial circadian clock of *Synechococcus 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)³¹. 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.
- SUN2 (**Figure 1F**), in complex with its binding partner KASH, tethers the nuclear envelope to the cytoskeleton, making it critical to meiosis, nuclear migration, and global cytoskeletal organization³². A region of SUN2, known as the KASH lid, is likely flexible in apo SUN2 and forms a small helix in the apo crystal structure. By contrast, the KASH lid forms a hydrogen-bonded β -sheet with KASH's C-terminal

segment in the complex (black). Both mutational and functional analyses highlight the importance of the SUN-KASH β -sheet: alanine mutations at conserved positions within the β -sheet of both SUN and KASH decreased binding significantly, and deletion of the KASH lid compromised targeting to the nuclear envelope³².

- The pathogenic malarial aspartic protease plasmepsin (a family of at least 10 closely related enzymes in *P. falciparum*³³; **Figure 1G**) degrades human hemoglobin and other globins – producing many of the disease’s symptoms³⁴. It has a zymogenic form with an N-terminal propeptide that targets it to the protist’s food vacuole, while also forcing the protein into a conformation that inhibits its activity. A maturase activates plasmepsin by cleaving its propeptide, the removal of which allows its N-terminus to dramatically refold into a conformation capable of proteolytic activity³⁴.
- Ovalbumin (**Figure 1H**) 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 protein³⁵) 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 serpins³⁶.
- The human amyloid-forming peptides amylin and amyloid β (**Figures I & J**, respectively), along with α -Synuclein (Table S1) are all believed to interact with membranes, where they form α -helices³⁷⁻³⁹. 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 metabolism³⁸. All three peptides can also form fibrillar deposits associated with diseases such as Parkinson's (α -synuclein)⁴⁰, type 2 diabetes (amylin)⁴¹, and Alzheimer's (amyloid β)⁴².

3.2. JPred4 predicts different secondary structures for isolated and contextualized FSRs

The proteins shown in **Figure 1** have fold-switching regions (FSRs, shown in color) and structurally constant regions (SCRs, shown in gray). Previous work suggests that the structures assumed by FSRs are context-dependent^{18,19}: that is, exogenous FSR-SCR interactions might bias the FSR away from its intrinsic secondary structure propensities, forcing it to adopt an alternative, higher-energy secondary structure conformation²⁰. Thus, we sought to determine whether JPred4 might predict that FSRs can fold into multiple stable secondary structures by comparing the secondary structure predictions of isolated FSR sequences (shown in **Figure 1**) with those of the FSR sequence within its natural sequence context (hereafter called contextualized FSRs).

In all cases shown above, along with all other KaiB variants tested (**Table S1**), we found that JPred4 predicted different secondary structures for isolated and contextualized FSR sequences (**Figure 1**). Furthermore, in 12/14 cases, the regions of α -helix $\leftrightarrow$ β -strand prediction discrepancy (where JPred4 predicted α -helix for an FSR sequence in one context and a β -strand for the same FSR in its other context) corresponded to a region where the sequence indeed assumes an α -helix in one experimentally determined conformation and a β -strand in the other. In the two cases where this did not occur (PfPRS and Ovalbumin, **Figure 1D and 1G**, respectively), one region of α -helix $\leftrightarrow$ β -strand prediction discrepancy was within 5 residues of an experimentally determined FSR that assumed an α -helix in one conformation and a β -strand in the other.

Figure 1. 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 do not. Predicted secondary structures that were at least 2 contiguous residues long are shown. The KaiB variant (2QKE) represents all members of the KaiB family; the other 3 (1WWJ, 4SKO, 1R5P) are not shown; α -Synuclein is also not shown due to lack of space. The C-terminus of KASH (SUN2 panel, green structure) is shown in black. All three-dimensional protein structures were made using PyMOL⁴⁸.

JPred4 secondary structure predictions tend to correspond reasonably well (<6% α -helix $\leftrightarrow$ β -strand discrepancies⁹) with at least one experimentally determined protein structure for all 14 proteins. In fact, in 10 of these 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, JPred4 predictions correspond well to both experimentally determined conformations in only 2/14 cases. Nevertheless, as in previous work¹³, 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's α -helix $\leftrightarrow$ β -strand prediction discrepancies are statistically significant for FSRs

We then sought to determine the significance of JPred4's α -helix $\leftrightarrow$ β -strand prediction discrepancies for isolated and contextualized FSRs. To do this, we randomly selected fragments from a set of 211 proteins expected not to switch folds (single-fold proteins). The resulting distribution of α -helix $\leftrightarrow$ β -strand discrepancies for fragments of single-fold proteins differed significantly from that of extant fold switchers ($p < 4.2 \times 10^{-20}$, Kolmogorov-Smirnov test; **Figure 2**).

We then used the fraction of α -helix $\leftrightarrow$ β -strand discrepancies as a potential classifier for predicting fold switchers. At a threshold of 0.5 (total number of helix $\leftrightarrow$ strand discrepancies normalized by total number of residues predicted to assume regular secondary structure, **Methods**), the Matthews Correlation Coefficient of this classifier was 0.47, demonstrating moderate agreement with experimental findings. With this constraint in

place, this predictive method yielded 9 true positives, 5 false negatives, 198 true negatives, and 13 false positives. Thus, we sought for additional constraints to improve predictions.

Figure 2. The distributions of JPred4 α -helix $\leftrightarrow$ β -strand prediction discrepancies between fragment:whole sequence for single-fold proteins and fold switchers differ significantly ($p < 4.2 \times 10^{-20}$, Kolmogorov-Smirnov test). Magenta dashed line represents the classifier threshold of 0.5, which yields a Matthews Correlation Coefficient of 0.47.

3.4. Constraining by fraction of secondary structure decreases false positives

Many false positives identified by the previous classifier shared a common feature: low overall levels of predicted secondary structure (and, conversely, high overall levels of predicted coil). To quantify the significance of this feature, we plotted the fraction of predicted secondary structure (total number of residues predicted to be helix or strand normalized by the total number of residues in the protein fragment) versus the fraction of secondary structure discrepancies for both FSRs and randomly selected regions of single-fold proteins (**Figure 3**). Indeed, we found that nearly all (12/13) false positives had low fractions of predicted secondary structure (< 0.35).

The one remaining false positive 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 family⁴³. 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 S2**)⁴³. Based on our previous work on sequence-similar fold switchers¹³, this result suggests that this protein segment might switch folds.

3.5. An improved classifier results from combining secondary structure prediction discrepancies with the fraction of secondary structure

Requiring a minimum of 35% secondary-structure content removed 12/13 false positives while eliminating only one true positive. Thus, we sought to classify fold switchers by thresholding our results by 0.5 for fraction of α -helix $\leftrightarrow$ β -strand discrepancies and 0.35 by fraction of predicted secondary structure (**Figure 3**). This yielded 8 true positives, 1 false positive, 210 true negatives, and 6 false negatives, resulting in a Matthews Correlation Coefficient of 0.70 (very low false positive rate; moderate false negative rate).

Figure 3. Predictions of extant fold switchers improve by constraining both the fraction of predicted secondary structure and the fraction of α -helix $\leftrightarrow$ β -strand discrepancies. Vertical/horizontal dashed lines represent the thresholds for fraction of predicted secondary structure/fraction of α -helix $\leftrightarrow$ β -strand discrepancies. Datapoints in the upper right quadrant are predicted to switch folds. Only 7/8 fold switchers can be seen because two KaiB variants have identical coordinates (0.37,0.69).

4. Discussion

Fold switchers are exceptions to the well-supported general 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 thought¹. Fold switching impacts protein function⁵ and is associated with a number of diseases^{2, 3, 30}. 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 chance¹.

Here we present an approach for predicting extant fold switchers from their amino acid sequences alone. To the best of our knowledge, this is the first sequence-based method for predicting extant fold switchers presented in the literature. This method is based on previous experimental work suggesting that the fold-switching regions (FSRs) of proteins are context-dependent: that is, their conformations are determined by their environment^{18, 19}. 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 $\leftrightarrow$ β -strand discrepancies. Indeed, such discrepancies were found in 14/14 fold switchers used in this study. Furthermore, this finding was highly statistically significant ($p < 4.2 \times 10^{-20}$, Kolmogorov-Smirnov test), and we used it to develop a classifier for extant fold switchers that yielded a Matthews Correlation Coefficient of 0.7. We suspect that JPred4 successfully identified extant fold switchers for the same reason it identified sequence-similar fold switchers¹³: 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 previous work. First of all, 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 variants¹³. 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 folds¹³). Additionally, Omega 3's GST topology⁴³ has been known to switch folds in other proteins, namely KaiB⁴⁴ and Chloride Intracellular Channel 1 (CLIC1)⁴⁵. Still, further experimental work would be needed to determine whether or not 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 $\leftrightarrow$ β -sheet transitions. To date, these proteins are rare and comprise only 15% of known fold switchers. Biologically important fold switchers similar to lymphotactin⁴⁶, which maintains β -sheets that change their hydrogen bonding register, and β -pore proteins⁴⁷, which extend already existing β -sheet structures, will be missed. Secondly, it will not identify all fold switchers that undergo large α -helix $\leftrightarrow$ β -sheet transitions, as evidenced by the fact that only 8/14 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 we probe, the more likely we are to hit false positives. Additional work will be needed to more accurately distinguish between these false positives and true fold switchers. Finally, our training set was small, comprising the only 14 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 $\leftrightarrow$ 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 $\leftrightarrow$ β -strand, its low false positive (1/211) and moderate true positive (8/14) 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.

Data availability

The data and code that support the findings of this study are openly available on GitHub at https://github.com/porterll/extant_fold_switchers

Acknowledgements

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 Statement

The authors declare no conflict of interest.

References

- Porter, L. L.; Looger, L. L., Extant fold-switching proteins are widespread. *Proc Natl Acad Sci U S A* **2018**, *115* (23), 5968-5973.
- Li, B. P.; Mao, Y. T.; Wang, Z.; Chen, Y. Y.; Wang, Y.; Zhai, C. Y.; Shi, B.; Liu, S. Y.; Liu, J. L.; Chen, J. Q., CLIC1 Promotes the Progression of Gastric Cancer by Regulating the MAPK/AKT Pathways. *Cell Physiol Biochem* **2018**, *46* (3), 907-924.
- Lei, Y.; Takahama, Y., XCL1 and XCR1 in the immune system. *Microbes Infect* **2012**, *14* (3), 262-7.
- Jain, V.; Kikuchi, H.; Oshima, Y.; Sharma, A.; Yogavel, M., Structural and functional analysis of the anti-malarial drug target prolyl-tRNA synthetase. *J Struct Funct Genomics* **2014**, *15* (4), 181-90.
- Kim, A. K., Porter, L.L., Functional and regulatory roles of fold-switching proteins. *Structure* **2020**, *In Press*.
- Chang, Y. G.; Cohen, S. E.; Phong, C.; Myers, W. K.; Kim, Y. I.; Tseng, R.; Lin, J.; Zhang, L.; Boyd, J. S.; Lee, Y.; Kang, S.; Lee, D.; Li, S.; Britt, R. D.; Rust, M. J.; Golden, S. S.; LiWang, A., Circadian rhythms. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science* **2015**, *349* (6245), 324-8.
- Kang, J. Y.; Mooney, R. A.; Nedialkov, Y.; Saba, J.; Mishanina, T. V.; Artsimovitch, I.; Landick, R.; Darst, S. A., Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **2018**, *173* (7), 1650-1662 e14.
- Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C., The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **2002**, *58* (Pt 6 No 1), 899-907.
- Mishra, S.; Looger, L. L.; Porter, L. L., Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci* **2019**, *28* (8), 1487-1493.
- Ambroggio, X. I.; Kuhlman, B., Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **2006**, *128* (4), 1154-61.
- Wei, K. Y.; Moschidi, D.; Bick, M. J.; Nerli, S.; McShan, A. C.; Carter, L. P.; Huang, P. S.; Fletcher, D. A.; Sgourakis, N. G.; Boyken, S. E.; Baker, D., Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proc Natl Acad Sci U S A* **2020**, *117* (13), 7208-7215.
- Tian, P., Best, R.B., Exploring the sequence fitness landscape of a bridge between two protein folds. *PLoS Comput Biol* **2020**, *16* (10), e1008285.
- Kim, A. K., Looger, L.L., Porter, L.L., A high-throughput predictive method for sequence-similar fold switchers. *Biopolymers* **2021**, *In press*.
- Sfriso, P.; Duran-Frigola, M.; Mosca, R.; Emperador, A.; Aloy, P.; Orozco, M., Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure* **2016**, *24* (1), 116-126.
- O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D., Reference sequence

(RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **2016**, *44* (D1), D733-45.

16. Gibrat, J. F.; Madej, T.; Bryant, S. H., Surprising similarities in structure comparison. *Curr Opin Struct Biol* **1996**, *6* (3), 377-85.
17. Huang, I. K.; Pei, J.; Grishin, N. V., Defining and predicting structurally conserved regions in protein superfamilies. *Bioinformatics* **2013**, *29* (2), 175-81.
18. Minor, D. L., Jr.; Kim, P. S., Context-dependent secondary structure formation of a designed protein sequence. *Nature* **1996**, *380* (6576), 730-4.
19. Porter, L. L.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N., Subdomain interactions foster the design of two protein pairs with approximately 80% sequence identity but different folds. *Biophys J* **2015**, *108* (1), 154-62.
20. Burmann, B. M.; Knauer, S. H.; Sevostyanova, A.; Schweimer, K.; Mooney, R. A.; Landick, R.; Artsimovitch, I.; Rosch, P., An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **2012**, *150* (2), 291-303.
21. Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J., JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* **2015**, *43* (W1), W389-94.
22. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577-637.
23. Holt, M. E.; Salay, L. E.; O'Brien, E.; Barton, J. K.; Chazin, W. J., Functional and structural similarity of human DNA primase [4Fe4S] cluster domain constructs. *PLoS One* **2018**, *13* (12), e0209345.
24. Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422-3.
25. Hunter, J. D., Matplotlib: A 2D graphics environment. *Comput Sci Eng* **2007**, *9* (3), 90-95.
26. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **1975**, *405* (2), 442-51.
27. Lopez-Pelegrin, M.; Cerda-Costa, N.; Cintas-Pedrola, A.; Herranz-Trillo, F.; Bernado, P.; Peinado, J. R.; Arolas, J. L.; Gomis-Ruth, F. X., Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metalloproteinase. *Angew Chem Int Ed Engl* **2014**, *53* (40), 10624-30.
28. Zuber, P. K.; Schweimer, K.; Rosch, P.; Artsimovitch, I.; Knauer, S. H., Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun* **2019**, *10* (1), 702.
29. Giganti, D.; Albesa-Jove, D.; Urresti, S.; Rodrigo-Unzueta, A.; Martinez, M. A.; Comino, N.; Barilone, N.; Bellinzoni, M.; Chenal, A.; Guerin, M. E.; Alzari, P. M., Secondary structure reshuffling modulates glycosyltransferase function at the membrane. *Nat Chem Biol* **2015**, *11* (1), 16-8.
30. Jain, V.; Yogavel, M.; Oshima, Y.; Kikuchi, H.; Touquet, B.; Hakimi, M. A.; Sharma, A., Structure of Prolyl-tRNA Synthetase-Halofuginone Complex Provides Basis for Development of Drugs against Malaria and Toxoplasmosis. *Structure* **2015**, *23* (5), 819-829.
31. Partch, C. L., Orchestration of Circadian Timing by Macromolecular Protein Assemblies. *J Mol Biol* **2020**, *432* (12), 3426-3448.
32. Sosa, B. A.; Rothballe, A.; Kutay, U.; Schwartz, T. U., LINC complexes form by binding of three KASH peptides to domain interfaces of trimeric SUN proteins. *Cell* **2012**, *149* (5), 1035-47.
33. Coombs, G. H.; Goldberg, D. E.; Klemba, M.; Berry, C.; Kay, J.; Mottram, J. C., Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends Parasitol* **2001**, *17* (11), 532-7.

34. Bernstein, N. K.; Cherney, M. M.; Yowell, C. A.; Dame, J. B.; James, M. N., Structural insights into the activation of *P. vivax* plasmepsin. *J Mol Biol* **2003**, *329* (3), 505-24.
35. Stein, P. E.; Leslie, A. G.; Finch, J. T.; Carrell, R. W., Crystal structure of uncleaved ovalbumin at 1.95 Å resolution. *J Mol Biol* **1991**, *221* (3), 941-59.
36. Yamasaki, M.; Arii, Y.; Mikami, B.; Hirose, M., Loop-inserted and thermostabilized structure of P1-P1' cleaved ovalbumin mutant R339T. *J Mol Biol* **2002**, *315* (2), 113-20.
37. Crescenzi, O.; Tomaselli, S.; Guerrini, R.; Salvadori, S.; D'Ursi, A. M.; Temussi, P. A.; Picone, D., Solution structure of the Alzheimer amyloid beta-peptide (1-42) in an apolar microenvironment. Similarity with a virus fusion domain. *Eur J Biochem* **2002**, *269* (22), 5642-8.
38. Patil, S. M.; Xu, S.; Sheftic, S. R.; Alexandrescu, A. T., Dynamic alpha-helix structure of micelle-bound human amylin. *J Biol Chem* **2009**, *284* (18), 11982-91.
39. Rao, J. N.; Jao, C. C.; Hegde, B. G.; Langen, R.; Ulmer, T. S., A combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. *J Am Chem Soc* **2010**, *132* (25), 8657-68.
40. Walti, M. A.; Ravotti, F.; Arai, H.; Glabe, C. G.; Wall, J. S.; Bockmann, A.; Guntert, P.; Meier, B. H.; Riek, R., Atomic-resolution structure of a disease-relevant Abeta(1-42) amyloid fibril. *Proc Natl Acad Sci U S A* **2016**, *113* (34), E4976-84.
41. Cao, Q.; Boyer, D. R.; Sawaya, M. R.; Ge, P.; Eisenberg, D. S., Cryo-EM structure and inhibitor design of human IAPP (amylin) fibrils. *Nat Struct Mol Biol* **2020**, *27* (7), 653-659.
42. Tuttle, M. D.; Comellas, G.; Nieuwkoop, A. J.; Covell, D. J.; Berthold, D. A.; Kloepper, K. D.; Courtney, J. M.; Kim, J. K.; Barclay, A. M.; Kendall, A.; Wan, W.; Stubbs, G.; Schwieters, C. D.; Lee, V. M.; George, J. M.; Rienstra, C. M., Solid-state NMR structure of a pathogenic fibril of full-length human alpha-synuclein. *Nat Struct Mol Biol* **2016**, *23* (5), 409-15.
43. Chen, B. Y.; Ma, X. X.; Guo, P. C.; Tan, X.; Li, W. F.; Yang, J. P.; Zhang, N. N.; Chen, Y.; Xia, Q.; Zhou, C. Z., Structure-guided activity restoration of the silkworm glutathione transferase Omega GSTO3-3. *J Mol Biol* **2011**, *412* (2), 204-11.
44. Tseng, R.; Goularte, N. F.; Chavan, A.; Luu, J.; Cohen, S. E.; Chang, Y. G.; Heisler, J.; Li, S.; Michael, A. K.; Tripathi, S.; Golden, S. S.; LiWang, A.; Partch, C. L., Structural basis of the day-night transition in a bacterial circadian clock. *Science* **2017**, *355* (6330), 1174-1180.
45. Littler, D. R.; Harrop, S. J.; Fairlie, W. D.; Brown, L. J.; Pankhurst, G. J.; Pankhurst, S.; DeMaere, M. Z.; Campbell, T. J.; Bauskin, A. R.; Tonini, R.; Mazzanti, M.; Breit, S. N.; Curmi, P. M., The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem* **2004**, *279* (10), 9298-305.
46. Dishman, A. F.; Tyler, R. C.; Fox, J. C.; Kleist, A. B.; Prehoda, K. E.; Babu, M. M.; Peterson, F. C.; Volkman, B. F., Evolution of fold switching in a metamorphic protein. *Science* **2021**, *371* (6524), 86-90.
47. Podobnik, M.; Savory, P.; Rojko, N.; Kisovec, M.; Wood, N.; Hambley, R.; Pugh, J.; Wallace, E. J.; McNeill, L.; Bruce, M.; Liko, I.; Allison, T. M.; Mehmood, S.; Yilmaz, N.; Kobayashi, T.; Gilbert, R. J.; Robinson, C. V.; Jayasinghe, L.; Anderluh, G., Crystal structure of an invertebrate cytolysin pore reveals unique properties and mechanism of assembly. *Nat Commun* **2016**, *7*, 11598.
48. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

Table S1: JPred4 predictions of extant fold switchers

4QHF/4QHH (Selecase)	HLKYGKYHINEFEEELLFLFPNKEA HHE-----HHHHHHHHHH--H -----EE--HHEEEEE-- (Fragment)
1MIQ/1QS8 (Plasmepsin)	LGSENDVIELDDVANIMFYGEDEVGDNHQKFM -----EEEE-----EEEEEE-----EEE -----EEEEHHHHHHHH-- (Fragment)
1OVA/1JTI (Ovalbumin)	LKISQAVHAAHAEINEAGREVVGSAEAGVDAASVSEEF -EEEEEEEEEEEE-----EEEE--EEE-----EE --HHHHHHHHHHHHHH--HHHH-- (Fragment)
4NC9/4N9W (PimA)	AAEGPIVATFHTSTTKSLTSLVF H---EEEE-----HHH -----EEEE-----EEEE-- (Fragment)
4DXT/4DXR (SUN)	TYETKTALLSLFGIPLWYHS -----EEEE----- ----HHHHHHHH--HHHH-- (Fragment)
2NAO/1IYT (Amyloid β)	DSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA ---HHHHH-----EEEEHHH---E-- ----EEE--EEEE-----EEEHHH---E-- (Fragment)
2N0A/2KKW (α -Synuclein)	EGVLYVGSKTKEGVVHGVATV...KEQVTNVGGAVVTGVTAVAQKTVE HHHHHEHH--HHHHHHHHHH...HHHHHHHHHHHHHHHHHHHHHHHHHHHH --EEEE-----EEEE--...--EE--EEEEEEHHH----- (Fragment)
6VW2/2KB8 (Amylin)	NFLVHSSNFGAILSSSTNVGSNTY HHHHH----- --EE-----EEE-----
4TWA/4YDQ (PfPRS)	NHIEGFSPEVAWVTKYGDSNLPEEIAIRPTSETIMYSVF -----EEEE-----EEEE-----EEEE -----EEEE-----EEE--HHHHHH-- (Fragment)
4KSO (KaiB)	LAKVLPLPVRRIIGDLSREKVLIGLDLLYGELQD EE-----EEEE--HHHHHHHH----- -----HHHHHH-----EEEEEE----- (Fragment)
1WWJ (KaiB)	LAKILPPPVRKIIGDLSREKVLIGLDLLYDEIRE EEE-----EEEE--HHHHHHHH----- -----HHHHHH-----HHHEEE----- (Fragment)

2QKE/5JYT (KaiB)	LAKVLPPPVRRIIGDLSNREKVLIGLDLLYEEIGD EEE----- EEEEEE -----HHHHHHHH----- -----HHHHHH----- EEEEEE ----- (Fragment)
1R5P (KaiB)	LSKILPPPVRKIIGDLSRERVLIIGLDLLYEELTE EEE----- EEEEEE -----HHHHHHHH----- -----HHHHHH-----HHHH EEE ----- (Fragment)
2LCL/2OUG (RfaH)	KVIITEGAFEGFQAI FTEPDGEARSMLLLNLINKEIKHSVKNTFRKL EEEE----- EEEEEE ----- EEEEEE -----EEEE----- EEEE----- HHHHH -----HHHHHHHHHH-----EEEE-----EE--

Table S2.

Sequence	I PPPALTDKLRLYHVD M NPYGHRVLLVLEAKRIKYEV*
3RBTA (FSR)	-----EEEEEE----- EEEEEE -----
3RBTA (whole)	-----EEEE-----HHHHHHHHHH-----EE
3RBTA (FSR N29C)	-----EEEEEE-----HHHHHHHHHH-----

*Position 29 (number from whole sequence, not the FSR) is shown in red.