ColabFold predicts alternative protein structures from single sequences, coevolution unnecessary for AF-cluster

Devlina Chakravarty[1], Joseph W. Schafer[1], Ethan A. Chen[1], and Lauren L. Porter[1,2,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894
[2]Biochemistry and Biophysics Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, 20892

*Correspondence: porterll@nih.gov

Abstract

Though typically associated with a single folded state, globular proteins are dynamic and often assume alternative or transient structures important for their functions[1,2]. Wayment-Steele, et al. steered ColabFold[3] to predict alternative structures of several proteins using a method they call AF-cluster[4]. They propose that AF-cluster "enables ColabFold to sample alternate states of known metamorphic proteins with high confidence" by first clustering multiple sequence alignments (MSAs) in a way that "deconvolves" coevolutionary information specific to different conformations and then using these clusters as input for ColabFold. Contrary to this Coevolutionary Assumption, clustered MSAs are not needed to make these predictions. Rather, these alternative structures can be predicted from single sequences, indicating that coevolutionary information is unnecessary for predictive success and may not be used at all.

We inputted single sequences of each successfully predicted target into ColabFold[3] and observed striking similarities between single-sequence and AF-cluster predictions (**Figure 1**). All predictions were performed on fold-switching proteins, which remodel their secondary and tertiary structures and modulate their functions in response to cellular stimuli[5]. Wayment-Steele and colleagues focused most of their predictions on homologs of KaiB, whose fold switching helps to regulate cyanobacterial circadian rhythm[6]. On average, single-sequence predictions of KaiB homologs were within 1.9 Å root-mean-squared deviation (RMSD) of their respective AF-cluster predictions. The fold-switching C-terminal domain (CTD) of the transcriptional regulator RfaH[7] predicted from a single sequence was within 1.8 Å RMSD of its corresponding AF-cluster prediction, and the C-terminal β-sheets of the mitotic spindle protein Mad2[8], though less accurate than the other structures (6.7Å), are poised in the alternative open conformation, on the left side of the 4-β-sheet rather than the right, which corresponds to the closed dominant state. Single-sequence predictions do not result from MSA-based coevolutionary inference, casting doubt on the claim that AF-cluster leverages coevolutionary information from its shallow MSAs containing 11 sequences including the target. These results are consistent with our previous analysis of Wayment-Steele et al.'s target Mpt53, in which we found no coevolutionary evidence for its putative alternative fold[9]. This prediction currently has no experimental support. If it is found to be correct, MSA-based coevolutionary inference is likely not the explanation.

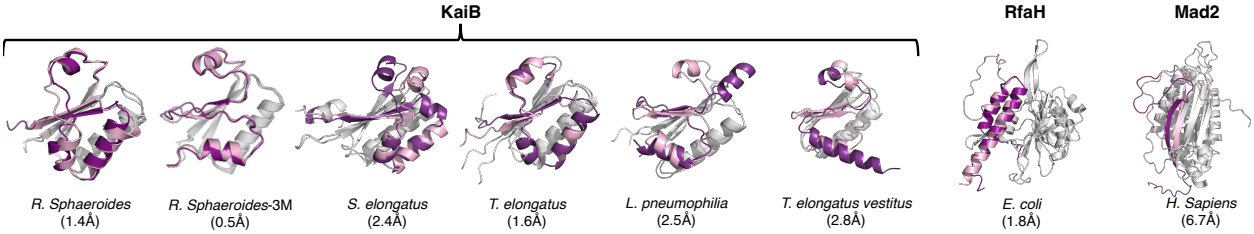

Figure 1. ColabFold predictions from single sequences match AF-cluster predictions. Superpositions of AF-cluster predictions (fold-switching regions in pink) with single-sequence predictions (fold-switching regions in purple); single-folding regions of both predictions colored light gray. All atom root mean squared deviations (RMSDs) reported below the source organisms of each protein.

Wayment-Steele et al. mentioned that single-sequence predictions of KaiB from *T. elongatus vestitus* (KaiB TV-4) were not consistent with its experimentally observed prediction[4]. We found that it was in 1/5 cases. Interestingly, 6/10 of the sequences in its cluster were predicted to assume the correct TV-4 conformation from single sequences. This suggested that ColabFold may use the structural inferences of these sequences to inform predictions of TV-4's structure more robustly, analogous to homology modeling rather than coevolutionary inference. To test this possibility, we ran ColabFold with an MSA of two sequences from this cluster with correct structure predictions. Two sequences are arguably too shallow for robust coevolutionary inference. Nevertheless, we found that this MSA boosted prediction quality considerably compared to single-sequence: 4/5 predictions were within 1.8Å of the AF-Cluster prediction. This result indicates that coevolution would not necessarily explain if AF-cluster's MSAs improve predictions, such as by requiring fewer recycling steps till the model converges. Proof of coevolution would need to rule out structural inference from sequence similarity, a difficult task with a complex neural network like ColabFold.

In their analysis, Wayment-Steele et al. used MSA Transformer[10] to confirm the presence of coevolutionary information in some of their shallow KaiB MSAs. A broader analysis revealed lack of substantial coevolutionary information in MSAs used in successful predictions. Here are three examples: one for the 2-sequence MSA of KaiB TV-4, which produces structures to within

1.0 Å of the 10-sequence AF-cluster prediction, one generated by AF-cluster and used to predict the structure of RfaH, and another generated by AF-cluster and used to predict the best structure of the alternative open form of Mad2. In all three cases, we see no evidence for coevolution unique to these structures (**Figure 2**), and in the case of RfaH, weak signals unique to the alternative β-sheet C-terminal domain are observed rather than the predicted α-helical conformation. MSA Transformer results are an estimate for the coevolutionary signals that ColabFold may detect. Nevertheless, Wayment-Steele et. al. use its couplings to argue for the presence of coevolutionary information in some of their MSAs, making it reasonable to use lack of couplings to argue for their absence. This is especially true since the authors of AlphaFold2[11]–the basis of ColabFold–state that consistent prediction accuracy requires an MSA of at least 32 input sequences: more than triple the size of AF-cluster's input MSAs.

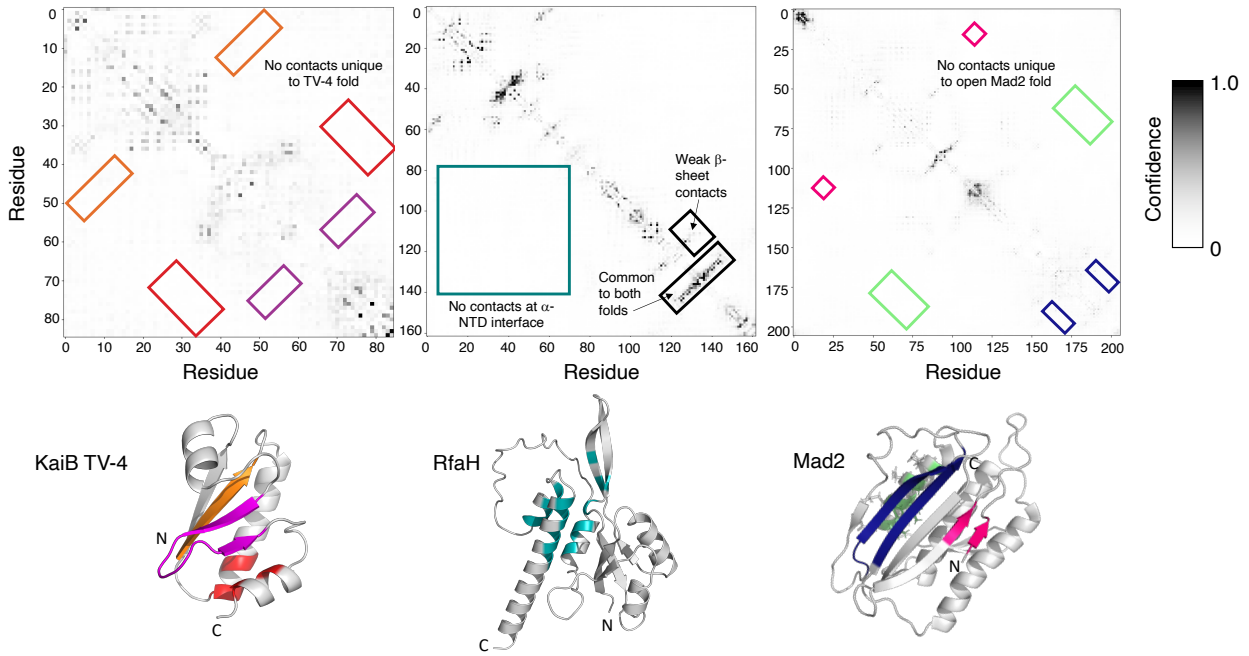

Figure 2. MSA Transformer predicts no unique coevolutionary signals from shallow MSAs used to predict the structures of KaiB TV-4, RfaH, or Mad2 successfully. Contacts unique to each fold are highlighted in their respective contact maps and colored according to their three-dimensional structures. Black boxes in the RfaH contact map are not represented in the structure and are used instead to annotate interesting features of its contact map.

If AF-cluster is not predicting these alternative conformations from coevolutionary information, what is the basis of its predictions? As a deep-learning algorithm, the specific inner-workings of ColabFold can be difficult to explain. However, we have reason to believe that ColabFold has memorized these alternative conformations and associates them with similar sequences, a possibility suggested by others[12].

AF-cluster fails to predict the alternative conformations of Selecase, Lymphotactin, and CLIC1. Nevertheless, our work has identified coevolutionary signals unique to all three of these missed conformations[9]. Interestingly, their ColabFold single-sequence predictions, unlike those of KaiB, RfaH, and Mad2, are biased toward their dominant states. One explanation is that Colabfold may not have stored information about their alternative conformations in its weights. This possibility may better explain Wayment-Steele et al.'s observation that AF-cluster fails to

predict oligomeric states of fold switchers, especially since it correctly predicts both the tetrameric KaiB fold and both Mad2 conformations, which dimerize.

We agree with Wayment-Steele and colleagues that coevolutionary information is present in shallow MSAs with sequences similar to the target of interest. In fact, we found direct coevolutionary evidence for the alternative conformations of 56/56 fold-switching proteins in shallow subfamily-specific MSAs–including all 6 families attempted by AF-cluster[9]. The issue is that there is no evidence that AF-cluster leverages this information to make its predictions, which are sometimes correct for other reasons. The next challenge for the field is to consistently predict both conformations of structurally diverse fold switchers–and eventually other alternative conformations–by leveraging this largely unused coevolutionary information. The idea behind AF-cluster suggests a reasonable way forward: retrain or fine-tune Colabfold to associate different conformations with MSAs directly shown to contain corresponding coevolutionary information.

Methods

Single sequence predictions

Single sequences of each AF-cluster prediction were inputted into ColabFold[3], the implementation of ColabFold used by AF-cluster, and run in single-sequence mode with all other settings default including 5 predictions per sequence. Predictions most similar (by RMSD) to each AF-cluster prediction–taken from the AF-cluster Github repo (www.github.com/HWaymentSteele/AF_Cluster)–are shown in Figure 1. Wayment-Steele et al. mentioned that single-sequence predictions of *T. elongatus vestus* were not consistent with its experimentally observed prediction[4]. We found that it was in 1/5 cases. Furthermore, running ColabFold with two-sequence MSAs (WP_011056401.1 and WP_058883586.1) improved predictions. These MSAs are too shallow for coevolutionary inference, and single-sequence predictions of WP_011056401.1 and WP_058883586.1–both taken from the *T. elongatus vestus* sequence cluster on the AF-cluster Github repo–yielded the correct *T. elongatus vestus* conformation.

Contact map generation

MSA Transfomer was run on a two-sequence MSA for KaiB TV-4 (WP_011056401.1 and WP_058883586.1), and the AF-cluster generated MSAs for RfaH (RFAH-049) and Mad2 (1S2H-047). Protein structures in both Figures were visualized with PyMOL[13].

Acknowledgements

L.L.P thanks Gabriel Rocklin for a helpful discussion. This work utilized resources from the NIH HPS Biowulf cluster (http://hpc.nih.gov), and it was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health (LM202011, L.L.P.).

References

1 Anthis, N. J. & Clore, G. M. Visualizing transient dark states by NMR spectroscopy. *Q Rev Biophys* **48**, 35-116, doi:10.1017/S0033583514000122 (2015).

2 Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508**, 331-339, doi:10.1038/nature13001 (2014).

3 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682, doi:10.1038/s41592-022-01488-1 (2022).

4 Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature*, doi:10.1038/s41586-023-06832-9 (2023).

5 Porter, L. L. & Looger, L. L. Extant fold-switching proteins are widespread. *Proc Natl Acad Sci U S A* **115**, 5968-5973, doi:10.1073/pnas.1800168115 (2018).
6 Chang, Y. G. *et al.* Circadian rhythms. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science* **349**, 324-328, doi:10.1126/science.1260031 (2015).
7 Burmann, B. M. *et al.* An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **150**, 291-303, doi:10.1016/j.cell.2012.05.042 (2012).
8 Mapelli, M., Massimiliano, L., Santaguida, S. & Musacchio, A. The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint. *Cell* **131**, 730-743, doi:10.1016/j.cell.2007.08.049 (2007).
9 Schafer, J. W. & Porter, L. Evolutionary selection of proteins with two folds. *Nat Commun* **14**, 5478 (2023).
10 Rao, R. M. *et al.* in *International Conference on Machine Learning*. 8844-8856 (PMLR).
11 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
12 Bryant, P. Structure prediction of alternative protein conformations. *bioRxiv*, 2023.2009.2025.559256 (2023).
13 The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.