A sequence-based foldability score combined with AlphaFold2 predictions to disentangle the protein order/disorder continuum

Apolline Bruley[1], Tristan Bitard-Feildel[1], Isabelle Callebaut[1], and Elodie Duprat[1]

[1]Institut de Mineralogie de Physique des Materiaux et de Cosmochimie

August 2, 2022

Abstract

Order and disorder govern protein functions, but there is a great diversity in disorder, from regions that are – and stay – fully disordered to conditional order. This diversity is still difficult to decipher even though it is encoded in the amino acid sequences. Here, we developed an analytic Python package, named *pyHCA*, to estimate the foldability of a protein segment from the only information of its amino acid sequence and based on a measure of its density in regular secondary structures associated with hydrophobic clusters, as defined by the Hydrophobic Cluster Analysis (HCA) approach. The tool was designed by optimizing the separation between foldable segments from databases of disorder (DisProt) and order (SCOPe (soluble domains) and OPM (transmembrane domains)). It allows to specify the ratio between order, embodied by regular secondary structures (either participating in the hydrophobic core of well-folded 3D structures or conditionally formed in intrinsically disordered regions) and disorder. We illustrated the relevance of *pyHCA* with several examples and applied it to the sequences of the proteomes of 21 species ranging from prokaryotes and archaea to unicellular and multicellular eukaryotes, for which structure models are provided in the AlphaFold2 databases. Cases of low-confidence scores related to disorder were distinguished from those of sequences that we identified as foldable but are still excluded from accurate modeling by AlphaFold2 due to a lack of sequence homologs or to compositional biases. Overall, our approach is complementary to AlphaFold2, providing guides to map structural innovations through evolutionary processes, at proteome and gene scales.

Hosted file

`Bruley_et_al_manuscript.docx` available at https://authorea.com/users/498784/articles/579495-a-sequence-based-foldability-score-combined-with-alphafold2-predictions-to-disentangle-the-protein-order-disorder-continuum