Title

Functional Prediction of Uncharacterized Proteins in Purple Sulfur Bacteria Using Structure, Sequence, and Homology Approaches.

Authors

Roxana Chavez[1] and Jonathon Mohl[1,2,3].

[1]Bioinformatics program, [2]Department of Mathematical Sciences, [3]Border Biomedical Research Center, The University of Texas at El Paso, El Paso, TX

Abstract

Many proteins in poorly studied organisms lack functional annotation because they share little sequence similarity with characterized proteins of more studied organisms. In this research, we evaluated whether protein structure prediction can help assign putative functions to uncharacterized proteins from purple sulfur bacteria (PSB). *Allochromatium vinosum* is a type of bacteria found in illuminated zones but absent of oxygen in lakes and ponds. They perform anoxygenic photosynthesis, using elemental sulfur as their energy source. It is highly important to understand the photosynthetic pathways used by purple sulfur bacteria (PSB), as they serve as model organisms for studying microbial energy conservation in harsh environments. The structures of 10 uncharacterized proteins were predicted using AlphaFold and I-TASSER, which are publicly available computational tools used to predict the 3D structure of proteins based on their amino acid sequence. The structural accuracy was measured by a confidence model using the pLDDT (predicted Local Distance Difference Test) score from AlphaFold and C-scores (Confidence score) from I-TASSER. The results between methods were visualized and tested in ChimeraX, an interactive visualization software with a matchmaker function that calculates the TM-score (Template Modeling) and RMSD (Root Mean Square Deviation). Putative functions were assigned from I-TASSER Gene Ontology predictions and structure similarity comparisons with sequence-based searches with BLAST and InterProScan results. All 10 proteins showed consistent folds between methods (TM-score > 0.5). However, InterProScan assigned a putative function to only one of them: Alvin_2092, which was classified as a sulfotransferase involved in carbohydrate biosynthesis. The rest of the proteins showed similarity with other proteins only found in other species of proteobacteria. Suggesting that the current methods are not effective characterizing bacterial genes and new approaches should be used. These results suggest that a combination of structure-base, sequence-based and many bioinformatic approaches can help annotate unknown proteins of understudied organisms.