

## ABSTRACT

Protein homodimers play a critical role in catalysis and regulation and their mechanism of folding is intriguing. The mechanisms of homodimer folding (2-state [2S] without intermediates and 3-state [3S] with either monomer [3SMI] or dimer [3SDI] intermediates) have been observed and documented for about 46 homodimers (27 2S; 12 3SMI; 7 3SDI) with known 3D structures. Determination of folding mechanisms through classical denaturation experiments is both time consuming, tedious, and expensive. Therefore, it is of interest to predict their folding mechanism. Furthermore, a large number of homodimers structures with unknown folding mechanism are available in the PDB. Hence, it is compelling to predict their folding mechanism using structural features intrinsic of each complex structure. Thus, we developed a classification and regression tree (CART) model using predictive parameters ((a) monomer protein size (ML); (b) interface area (B/2); (c) interface to total residues (I/T) ratio) derived from a dataset (46 homodimers with both known structures and folding mechanism) for folding mechanisms prediction. The dataset was subjectively divided into training (13 2S; 6 3SMI; 3 3SDI) and testing (14 2S; 6 3SMI; 4 3SDI) sets for validation. The model performed well for predicting 2S and 3SMI in both during training and testing using ML and I/T as predictive variables. However, it should be noted that the performance of model in classifying 3SDI is weak. Nonetheless, the model was not stable with the inclusion of the predictive variable B/2 and hence, was not considered during training and testing. The CART model produced accuracies of 85% (2S), 83% (3SMI) and 100% (3SDI) with positive predictive values (PPV) of 100% (2S), 83% (3SMI) and 75% (3SDI) during training. It then produced accuracies of 100% (2S) and 50% (3SMI) with positive predictive values (PPV) of 74% (2S), 60% (3SMI) during testing. Thus, we then used the model to assign

folding mechanisms to protein homodimers with known structures and unknown folding mechanisms. This exercise provides a framework for predicted homodimer structures with unknown folding mechanism for further verification through folding experiments. The CART model was able to assign folding mechanisms to all (169) the homodimer structures (with unknown folding data) due its automatically robust learning capabilities unlike the decision model with potential un-assignments.