Abstract

The arrangement of amino acids in a protein sequence encodes its native folding. However, the same arrangement in aggregation-prone regions may cause misfolding as a result of local environmental stress. Under normal interior to avoid aggregation and attain the native fold. We have used solvent accessibility of aggregation patches (SAAPp) to determine the packing of aggregation-prone residues. Our results showed that SAAPp has low values for native crystal structures, consistent with protein folding as a mechanism to minimize the solvent accessibility of aggregation-prone residues. SAAPp also shows an average correlation of 0.76 with the global distance test (GDT) score on CASP12 template-based protein models. Using SAAPp scores and five structural features, a random forest machine learning quality assessment tool, SAAP-QA, showed 2.32 average GDT loss between best model predicted and actual best based on GDT score on independent CASP test data, with the ability to discriminate native-like folds having an AUC of 0.94. Overall, the Pearson correlation coefficient (PCC) between true and predicted GDT scores on independent CASP data was 0.86 while on the external CAMEO dataset, comprising high quality protein structures, PCC and average GDT loss were 0.71 and 4.46 respectively. SAAP-QA can be used to detect the quality of models and iteratively improve them to native or near-native structures.