
Protein structure classification and comparison has become a central area in the field of bioinformatics. The rapid increase in the size of protein databases has prompted the development of rapid, automated methods to classify unknown protein structures. Protein structural databases commonly suffer from the 'curse of dimensionality,' necessitating the development of novel dimensionality reduction of protein structural information prior to classification. Moreover, the design and development of efficient manual or semi-automated classification techniques have not kept pace with the growth in such databases. In this paper, we propose a novel, automated computational framework for the three dimensional (3D) structure-based classification of proteins using an orthogonal transformation of geometric shape descriptors derived from protein structures by employing an association rule-based, supervised clustering approach to classify proteins.

This research incorporates two previously proposed structural descriptors, dihedral angle and bond length, to represent the 3D protein structure. The distributions of these descriptors over a sequence are then orthogonally transformed into corresponding signals in the frequency domain using DCT, followed by selective feature filtering. Associations between the coefficients produced by the DCT process are used to derive classes that represent a particular protein structure. Class-association rule discovery is used to identify such associations in a group of proteins that belong to a structural class. To demonstrate the sensitivity and specificity of the approach, we employ our method to two different datasets. The first balanced dataset consisted of 400 proteins from 10 families. The 3D protein structure information was extracted from the PDB files, referred family-wise from the SCOP database. We experimented with 1D and 2D DCT and found that higher classification accuracy (over 85%) was attained for 2D DCT. In our second experiment, we implemented our framework on a dataset of 600 proteins from 15 folds. Our method demonstrated an overall accuracy of better than 83%. Thus, the proposed novel computational framework demonstrates the applicability of rule discovery-based classification of structural descriptors for protein fold classification with improved sensitivity.