Abstract

Protein function prediction is an important and challenging field in Bioinformatics. There are various machine learning based approaches have been proposed to predict the protein functions using sequence derived properties. In this paper 857 sequence-derived features such as amino acid composition, dipeptide composition, correlation, composition, transition and distribution and pseudo amino acid composition are used with various machine learning based approaches such as Random Forest, Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), and fuzzy k-Nearest Neighbor (k-NN) to predict the protein functions. This paper used various feature selection techniques such as Correlation Feature Selection, Gain Ratio, Information Gain, One R attribute, ReliefF to select the optimal number of features. The performance of various classifiers with optimal number of features obtained by various feature selection techniques. The comparative analysis of result shows that the random forest based method with reliefF provide the overall accuracy of 89.20% and Matthews' s correlation coefficient (MCC) 0.87% that is better to others.
References


Index Terms

Computer Science

Information Science

Keywords

Protein function Classification Random Forest SVM k-NN fuzzy k-NN
Machine Learning based Approach for protein Function Prediction using Sequence Derived Properties