Adaptive Resonance Theory-2 Neural Network for Protein Classification

U. B. Angadi and M. Venkatesulu

Department of Computer Applications,
Kalasalingam University, Krishnankoil,
Srivilliputtur (via), Tamil Nadu, India, 626190.
angadiub@gmail.com

Clustering of proteins is a major direction for defining the relationships in terms of sequences structure and functions. These relationships are embodied in hierarchical structure such as structural classification of proteins (SCOP) and CATH (Class, Architecture, Topology and Homology). These employ different heuristic, similarity measures and different degree of automation. SCOP is manually curated and which is well accepted by research as most accurate structural classification. Such complex classification is essential to structural and functional analysis of proteins and large number of proteins remain listed under the not a true class. Work on protein structural classification has been ongoing for over a decade and many methods are currently available for clustering of protein sequences into folds, superfamilies and family. They have difficulties in classifying protein superfamily sharing very weak similarity. In such case these methods fail to provide correct information about the classification of the sequence. A supervised machine learning algorithm classifiers show better results, but these approaches are based on what it learns from classification already available in SCOP and not only on pair wise sequences alignment and similarity measures. Some methods use various sequence comparison such as BLAST, PSIBLAST, COMPASS, DALI and MAMMOTH and classified into superfamily using hard threshold.

We propose unsupervised machine learning ART2 neural network algorithm to classify a given set proteins into SCOP superfamilies. In the proposed method, we construct a similarity matrix from P-values of BLAST all-against-all, train the network with ART-2 unsupervised-learning algorithm using similarity matrix as input vectors and finally the trained network offers SCOP super-family level classification. We use the three data set of 507 sequences , 457 sequences and 342 sequences. These sets contains domains from the Globin-like proteins, EF-hand proteins, Cuperdoxins proteins, (Trans) glycosidases proteins, Thioredoxin-like proteins and Membrane all-alpha of SCOP We have chosen two kinds of experiments. In first, we trained the network with randomly shuffled dataset individually as training dataset and the same dataset used as test data. In second experiment, we have trained the network with randomly shuffled 457 sequences dataset as training dataset and tested with the other two 507 and 342 sequences datasets. Results are variant in random shuffling of training dataset, varies from 85% to 97% accuracy for 342 dataset, from 82% to 95% for 457 dataset, and from 80% to 96% for 507 dataset. The trained network is able to classify a given similarity matrix of a set of sequences at up to 97% f-measure accuracy.