An Improved Deep Learning Based Method for Protein Family Classification

P.D. Sandaruwan* and C.T. Wannige

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka.

Proteins are large and complex molecules that play a critical role in various aspects of life. Only 20 amino acids provide millions of proteins by combining into chains called polypeptide chains with different types of amino acids, lengths, and folds. Therefore, they are considered the building blocks of life. In proteomics, proteins are classified into families to achieve many goals such as predicting functional properties of novel proteins, discovering new drugs for new diseases, etc. As biological experiments are more expensive to deal with a large number of new proteins, one of the main computational approaches of protein classification is deep learning. Nowadays, with the progress of computational techniques, deep learning plays a key role in many areas. In this paper, our goal is to offer an improved alignment-free deep learning-based method for pattern recognition in proteins for classification. In this research work, we were based on one of the recent deep Learning approaches called DeepFam. We designed an improved method using the concepts which have been used in image classification and natural language processing. We extensively experimented with using the Clusters of orthologous Groups (COG) and G-Protein-coupled receptor (GPCR) datasets. our method showed higher validation accuracy than DeepFam and other methods that had been experimented using the same data sets.

Keywords: Bioinformatics, Protein family prediction, Deep learning, Deep convolutional neural networks