Mr. Maqsood Hayat

Ph.D. SCHOLAR, DCIS PIEAS

Abstract

Membrane proteins are the basic constituent of a cell that manage intra and extracellular processes of a cell. About 20-30% of genes of eukaryotic organisms are encoded from membrane proteins. In addition, almost 50% of drugs are directly targeted against membrane proteins. Owing to the significant role of membrane proteins in living organisms, the identification of membrane proteins with substantial accuracy is essential. However, the annotation of membrane proteins through conventional methods is difficult, sometimes even impossible. Therefore, membrane proteins are predicted from topogenic sequences using computational intelligence techniques. In this study, we conducted our research in two phases regarding the prediction of membrane protein types and structures. In Phase-I, regarding the prediction of membrane protein types, four different ways are explored in order to enhance true prediction.

In the first part of phase-I, membrane protein types are predicted using Composite protein sequence representation followed by the application of principal component analysis in conjunction with individual classifiers. In the second part, the notion of ensemble classification is implemented by two different ways: simple majority voting and genetic algorithm based ensemble classification. Discrete wavelet analysis, Pseudo amino acid (PseAA) composition, and Split amino acid composition (SAAC) as well as hybrid models of entire space and reduced feature space are used to express membrane protein sequences. In part three, the performance of Support Vector Machine with and without error correction is investigated using evolutionary profiles (Position Specific Scoring Matrix) and SAAC based features. Finally, in part four, a

two-layer web predictor (Mem-PHybrid) is developed. In this model, protein sequences are represented by a hybrid model of physicochemical properties of amino acids and SAAC. In the sequel, Minimum redundancy and Maximum relevance feature reduction technique is employed. Mem-PHybrid accomplishes the prediction in two steps. First, a protein query is identified as a membrane or a non-membrane. In case of membrane protein, then its type is predicted.

In the second phase of this research, the structure of membrane protein is recognized as alpha-helix transmembrane or outer membrane proteins. In case of alpha-helix transmembrane proteins, protein sequences are expressed by two feature extraction schemes of distinct natures; including physicochemical properties and compositional indices of amino acids. Singular value decomposition is employed to extract high variation features, whereby, only five features are selected from each feature space and are then merged to form a hybrid model. Weighted Random Forest is used as a classification algorithm. On the other hand, in case of outer membrane proteins, protein sequences are represented by Amino acid composition, PseAA composition, and SAAC along with their hybrid models. Genetic programming, K-nearest neighbor, and fuzzy K-nearest neighbor are adopted as classification algorithms.

Through the simulation study, we observed that the prediction performance of our proposed approaches in case of both types and structures prediction is better compared to existing state of the arts/approaches. Finally, we conclude that our proposed approach for membrane proteins might play a significant role in Computational Biology, Molecular Biology, Bioinformatics, and this might be helpful in applications related to drug discovery. In addition, the related web predictors provide sufficient information to researchers and academicians in future research.