GPCR-PEn: A Database and Computational Framework for Prediction and Classification of G Protein-Coupled Receptors

Khodeza Begum,[1,2] Melanie Sosa,[2] Jonathon Mohl,[1,2,3] and Ming-Ying Leung[1,2,3]
[1]Bioinformatics Program, [2]Border Biomedical Research Center, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX 79968, USA
kbegum@utep.edu

Abstract

G protein-coupled receptors (GPCRs) are the largest family of membrane proteins, playing key roles in vision, olfaction, inflammation, and serving as major drug targets. As protein sequence data continues to grow, computational analysis of GPCRs is essential for understanding their biological functions and ligand-binding properties across a variety of species ranging from microinvertebrates, such as rotifers, to humans. In this work, we have developed a curated database and assembled computational algorithms for predicting and classifying GPCRs up to four hierarchical levels, namely family, sub-family, sub-sub-family, and sub-type. GPCR Prediction Ensemble (GPCR-PEn) provides over 7,000 protein sequences with comprehensive feature annotations that includes counts of amino acids, dipeptides, and region-specific lengths. Our pipeline also integrates multiple machine learning algorithms, with methods like support vector machines, hidden Markov models and neural networks, each optimized for distinct levels of GPCR prediction. By integrating predictive models with a curated sequence database, users can explore GPCRs to analyze their classification, structure, and potential functions. We also demonstrate the robustness of our computational models through extensive testing and their ability to efficiently handle large-scale datasets with high performance and scalability.