

## TRANSMEMBRANE HELIX AND TOPOLOGY PREDICTION USING HIERARCHICAL SVM CLASSIFIERS AND AN ALTERNATING GEOMETRIC SCORING FUNCTION

Allan Lo<sup>1,2</sup>, Hua-Sheng Chiu<sup>3</sup>, Ting-Yi Sung<sup>3</sup>, Wen-Lian Hsu<sup>3</sup>

<sup>1</sup>Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan

<sup>2</sup>Department of Life Sciences, National Tsing Hua University, Hsinchu, Taiwan

<sup>3</sup>Bioinformatics Lab., Institute of Information Science, Academia Sinica, Taipei, Taiwan

A key class of membrane proteins contains one or more transmembrane (TM) helices, traversing the membrane lipid bilayer. Various properties such as the length, arrangement and topology or orientation of TM helices, are closely related to a protein's functions. Although a range of methods have been developed to predict TM helices and their topologies, no single method consistently outperforms the others. In addition, topology prediction has much lower accuracy than helix prediction, and thus requires continuous improvements. Results: We develop a method based on support vector machines (SVM) in a hierarchical framework to predict TM helices first, followed by their topology. By partitioning the prediction problem into two steps, specific input features can be selected and integrated in each step. We also propose a novel scoring function for topology models based on membrane protein folding process. When benchmarked against other methods in terms of performance, our approach achieves the highest scores at 86% in helix prediction ( $Q_2$ ) and 91% in topology prediction (*TOPO*) for the high-resolution data set, resulting in an improvement of 6% and 14% in their respective categories over the second best method. Furthermore, we demonstrate the ability of our method to discriminate between membrane and non-membrane proteins, with higher than 99% in accuracy. When tested on a small set of newly solved structures of membrane proteins, our method overcomes some of the difficulties in predicting TM helices by incorporating multiple biological input features.