

Correlation of the performance of protein structure prediction algorithms with the size of the training set

¹Tsaousis G.N., ^{1,2}Bagos P.G. and ¹Hamodrakas S.J.

¹Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 157 01

²Department of Biomedical Informatics, School of Applied Sciences, University of Central Greece, Lamia, 35 100.

It is now evident, that the progress in the determination of membrane protein structure, grows exponentially, with approximately the same growth rate as that of the water-soluble proteins. In order to investigate the effect of this fact on the performance of the predictive algorithms for both alpha-helical and beta-barrel membrane proteins, we conducted a prospective study based in historical records. For this reason, we trained separate HMMs with different sized training sets and evaluated their performance on topology prediction for the two classes of transmembrane proteins. For model fitting we used either the non-linear model of von Bertalanffy or alternatively a linear regression model. We show that the existing top-scoring predictive algorithms of transmembrane segments of alpha-helical membrane proteins perform slightly better than those of beta-barrel outer membrane proteins (88% compared to 87%). With the same rationale, a meta-analysis of the performance of the secondary structure prediction algorithms indicates that existing algorithmic techniques cannot be further improved by just adding more non-homologous sequences to the training sets. This way, the upper limit for the secondary structure prediction is estimated to be no more than 70% and 80% of correctly predicted residues for single sequence based methods and multiple sequence alignment based ones, respectively. Furthermore, the results of this study suggest that we have reached a plateau, after which the predictive performance will not be further improved. Therefore, we should concentrate our efforts on utilizing new techniques for the development of even better scoring predictors. Differences in the estimated rates for water-soluble proteins, alpha-helical and beta-barrel membrane proteins with respect to the size of the training set used are highlighted, and potential implications for future studies are discussed.