Methodology, performance and retrainability survey of intrinsic disorder predictors

Intrinsically disordered proteins and regions are widely distributed within most proteomes. Recent studies show that they are associated with many essential biological processes and a broad range of human diseases. Given the prevalence of disordered proteins and the growing acknowledgement of their functional relevance, considerable effort has been made by the bioinformatics community to provide computational tools to predict protein disorder. To date, based on various characteristics of protein disorder, along with variety of diverse computational approaches, numerous disorder predictors have been developed.

Over the past decade several review papers examining intrinsic disorder predictors have been published. All these papers have played a significant role in stimulating and greatly facilitating the development of this actively growing field by pinpointing the potential room for improvement. Inspired by these, in this work we aim to integrate the relevant information regarding the existing intrinsic disorder predictors from the corresponding research papers in a novel review, including latest prediction tools. In addition, for each disorder predictor, we examined the possibility of their retraining using different datasets.

Here, we present an overview of 23 protein disorder prediction methods, including the thorough analysis of their advantages and weaknesses which derive from their different computational approaches. Regarding this, we precisely describe the methodology used for building the models and categorize them by different classification schemes. The performance of these models is presented by their scores from the most recent CAID competition. Additional contribution of this work is the models' retraining availability analysis. We describe in detail the predictors' implementation source code (if available) and propose a way around to overcome the obstacles with retraining procedure (if possible). This insight might be very useful, since older models were trained on significantly smaller datasets compared to the newer ones, due to the increase in the number of experimentally annotated disorder proteins with time. With respect to this, we discuss in detail the possibility of retraining the models on a different (bigger, novel) dataset in order to perform full-scale comparison of their expression power in delineating disorder in proteins.