High-throughput prediction of disordered flexible linker regions

Lukasz Kurgan[1] and Fanchi Meng[2]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, U.S.A.
[2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4, Canada

Abstract

Disordered flexible linkers (DFLs) are disordered regions that serve as flexible linkers/spacers in multi-domain proteins or disordered domains that link other domains. They are experimentally determined with X-ray crystallography, NMR and circular dichroism, and there is no computational method that can directly predict such regions from protein sequences. We conceptualized, developed and empirically assessed the first sequence-based predictor of DFLs, DFLpred. This method outputs propensity to form DFLs for each amino acid (AA) from an input protein sequence. DFLpred uses a small set of empirically selected features that quantify propensity towards formation of certain secondary structures, disordered regions and potential for inclusion in globular domains, which are processed by a fast logistic regression-based predictive model. Our high-throughput predictor can be used on the whole-proteome scale; it takes < 1 minute to predict 100 proteins with average size of 500 AAs on a single CPU. When assessed on an independent test dataset, DFLpred secures area under the ROC curve (AUC) equal 0.715 and outperforms alternatives that include methods for the prediction of flexible linkers, flexible residues (B-factors) and intrinsically disordered residues. Predictions on human proteome reveal that estimated 4.4% of human proteins have a large content of over 30% of residues in DFLs. We also estimate that most DFLs are short with only 2.9% that are 30 or more residues long and thus could form domains. DFLpred is available at http://biomine.ece.ualberta.ca/DFLpred/.