Propensity for X-ray crystallography and structural coverage of the protein universe

Lukasz Kurgan
Electrical and Computer Engineering
University of Alberta

Abstract
Structural Genomics (SG) is an international effort that aims at solving three-dimensional structures of important biological macro-molecules with primary focus on proteins. One of the main bottlenecks in SG is the ability to produce diffraction quality crystals for the X-ray crystallography-based protein structure determination. SG pipelines allow for certain flexibility in target selection which motivates development of in-silico methods for sequence-based prediction/assessment of the protein crystallization propensity.

We will overview the currently available sequence-based predictors of crystallization propensity, focusing on two of our recent methods: PPCpred and fDETECT. PPCpred alleviates drawbacks of the prior methods by using more recent data and improved protocol to annotate progress along the crystallization pipeline. This is the first predictor capable of predicting the success of the entire process (similar to the prior methods) and also several steps of the pipeline including production of crystals, purification, and production of the protein material. fDETECT is our newest predictor that provides similar predictive performance compared to PPCpred and substantially shorter runtime. Utilizing fDETECT, we answer the question whether three-dimensional structures of all protein families can be determined using X-ray crystallography? This is based on a first-of-its-kind analysis of crystallization propensity for a current snapshot of the protein universe consisting of all proteins encoded in 1,953 fully sequenced genomes across eukaryotes, bacteria, archaea, and viruses.