Structural coverage of protein universe

Lukasz Kurgan
Computer Science, Virginia Commonwealth University

Abstract

The already huge and rapidly expanding protein universe, defined as a set of all proteins of all organisms, creates a substantial challenge to understand and functionally annotate these bigdata. Structural Genomics (SG) is an international effort to solve structures of important biological macro-molecules, focusing primarily on mapping structures of the entire protein universe. Knowledge of protein structures is necessary to understand their biochemical and cellular functions, to decipher whether and how they interact with other molecules, including drugs.

One of the main bottlenecks in SG is a very low success rate to produce diffraction quality crystals for the X-ray crystallography, the dominant method for the determination of protein structures. SG pipelines allow for certain flexibility in selection of protein targets, and this motivates development of computational methods for the prediction/assessment of the protein crystallization propensity.

We will overview the currently available computational predictors of crystallization propensity. We will focus on our newest method, fDETECT, which provides good predictive performance coupled with short runtime. Utilizing fDETECT, we will answer the question whether the structures of all protein families can be determined with the help of the X-ray crystallography? We will summarize our first-of-its-kind analysis of crystallization propensity for a current snapshot of the protein universe consisting of over 8 million proteins encoded in 1953 fully sequenced genomes across eukaryotes, bacteria, archaea, and viruses. We will demonstrate that mapping of the protein universe is far from being complete but could be substantially improved with the X-ray crystallography.