Methods for protein folding rate prediction

Natalya S. Bogatyreva, Dmitry N. Ivankov

Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain; Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain; Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow Region, 142290, Russia, ivankov13@gmail.com

After revolutionary experiment of Anfincen, the main question in protein physics is how do proteins fold? The progress, both in experiments and theories, resulted in ability to simulate process of folding. Here we review the methods for prediction protein folding rates.

According to the underlying logic, the methods fall into four categories. Methods of the first group treat protein folding as a random search. The second group of the methods are based on the nucleation mechanism, discovered both experimentally and in computer simulations. The third group hypothesizes major role of protein topology and locality of native contacts on the folding rates. The fourth group consists of statistical and machine learning methods.

From the instrumental point of view, the methods can be divided into analytical theories, computer simulations, empirical models and statistical/machine learning methods. As input, the methods can take three-dimensional protein structure, secondary structure, amino acid sequence or just the number of amino acid residues comprising protein. Some methods are designed to predict folding rates for all proteins; the other are designated to specific structural classes (α-, β-, α/β, or α + β), or type of protein folding kinetics (with intermediates or not).

We noticed in our study that high correlation coefficient doesn't necessarily mean the correctness of the model from viewpoint of physics. We also noticed that statistical/machine learning methods often don't comply with the statistical protocols. Overall, the most adequate and robust methods are based on the nucleation mechanism.

The work has been supported by two grants from the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017 (Sev-2012-0208)' and (BFU2012-31329), the European Union and the European Research Council grant (335980_EinME), RFBR (13-04-00253a), MCB RAS (01201358029) and MES RK Grants.