Testing applicability of machine learning for protein folding rate prediction

Marc Corrales[1,2]*, Pol Cuscó[1,2]*, Dinara R. Usmanova[1,2]*, Heng-Chang Chen[1,2], Natalya S. Bogatyreva[1,2,3], Guillaume J. Filion[1,2], Dmitry N. Ivankov[1,2,3]

[1]*Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona,*

[2]*Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain,*

[3]*Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow Region, 142290, Russia*

Correspondence should be addressed: *ivankov13@gmail.com*

* Authors contributed equally to the work.

Understanding and predicting the self-organization of protein structure is one of the most important problems of the last 50 years in biophysics [1]. Massive experimental and theoretical efforts have led to a better understanding of protein folding [2], better prediction of protein structure [3–5] and successful *de novo* protein design [6]. In spite of this progress, apparently simple computational problems related to protein structure still remain challenging. Among them, predicting the rate of protein folding, *i.e.* the speed at which a protein renatures in the proper chemical environment, has shown only limited success. Yet, the ability to predict protein folding rates is instrumental for our understanding of the general principles of the protein folding process.

Due to the increasing amount of experimental data [7], numerous protein folding models and predictors of protein folding rates have been developed in the last decades. The problem has also attracted the attention of scientists from computational fields, which led to the publication of a flurry of machine learning-based models to predict the rate of protein folding. Several of these studies claim to predict the folding rate with an accuracy greater than 90% [8,9]. However, there are reasons to believe that such claims are exaggerated due to large fluctuations and overfitting of the estimates. When we confronted three example published models [8-10] with new data, we found a much lower predictive power than reported in the original publications. Based on this, we highlight common methodological mistakes in the studies claiming extravagant prediction power.

Additionally, we explored the particular class of linear models based on amino acid composition. The analysis of learning curve demonstrates clearly that there is presently not enough experimental data to properly train the complete model.

In summary, the lack of experimental data on the rate of protein folding is such that model fitting suffers large fluctuations, even for models with a single parameter. For more complex models, the situation is more severe because they can be overfitted easier. As a recommendation for future studies, we suggest to use learning curves to demonstrate the validity of the models instead of correlations and p-values.

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.

1. K.A.Dill, J.L.MacCallum (2012) The protein-folding problem, 50 years on. *Science*, **338**:1042–1046.

2. A.Sali, E.Shakhnovich, M.Karplus (1994) How does a protein fold? *Nature*, **369**:248–251.

3. J.Moult *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins*, **82 Suppl 2**:1–6.

4. T.A.Hopf *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**:1607–1621.

5. J.I.Sułkowska *et al.* (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci U S A*, **109**:10340–10345.

6. N.Koga *et al.* (2012) Principles for designing ideal protein structures. *Nature*, **491**:222–227.

7. N.S.Bogatyreva *et al.* (2009) KineticDB: a database of protein folding kinetics. *Nucleic Acids Res* **37**:D342–346.

8. M.M.Gromiha (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J Chem Inf Model*, **45**:494–501.

9. M.M.Gromiha *et al.* (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res*, **34**:W70–74.

10. Huang J-T, Tian J (2006) Amino acid sequence predicts folding rate for middle-size two-state proteins. *Proteins,* **63**:551–554.