PLOS | ONE

RESEARCH ARTICLE

Machine Learning: How Much Does It Tell about Protein Folding Rates?

Marc Corrales[1,2,3☯], **Pol Cuscó**[1,2,3☯], **Dinara R. Usmanova**[2,4,5☯], **Heng-Chang Chen**[1,2,3], **Natalya S. Bogatyreva**[2,4,6], **Guillaume J. Filion**[1,2,3], **Dmitry N. Ivankov**[2,4,6]*

1 Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain, **2** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **3** Spain Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain, **4** Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain, **5** Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia, **6** Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, Pushchino, Moscow Region, Russia

☯ These authors contributed equally to this work.
* ivankov13@gmail.com

Abstract

The prediction of protein folding rates is a necessary step towards understanding the principles of protein folding. Due to the increasing amount of experimental data, numerous protein folding models and predictors of protein folding rates have been developed in the last decade. The problem has also attracted the attention of scientists from computational fields, which led to the publication of several machine learning-based models to predict the rate of protein folding. Some of them claim to predict the logarithm of protein folding rate with an accuracy greater than 90%. However, there are reasons to believe that such claims are exaggerated due to large fluctuations and overfitting of the estimates. When we confronted three selected published models with new data, we found a much lower predictive power than reported in the original publications. Overly optimistic predictive powers appear from violations of the basic principles of machine-learning. We highlight common misconceptions in the studies claiming excessive predictive power and propose to use learning curves as a safeguard against those mistakes. As an example, we show that the current amount of experimental data is insufficient to build a linear predictor of logarithms of folding rates based on protein amino acid composition.

Introduction

Understanding 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] culminating in successful predictions of protein structures [3–6] and *de novo* protein design [7]. In the light of this spectacular progress, apparently simpler tasks still remain problematic. One of them is predicting the rate of protein folding, *i.e.* the speed at which a protein renatures *in vitro* in conditions matching the physiology. Surprisingly, proteins fold fast (from microseconds [8] to hours [9]) even though the number

OPEN ACCESS

Citation: Corrales M, Cuscó P, Usmanova DR, Chen H-C, Bogatyreva NS, Filion GJ, et al. (2015) Machine Learning: How Much Does It Tell about Protein Folding Rates? PLoS ONE 10(11): e0143166. doi:10.1371/journal.pone.0143166

Editor: Emanuele Paci, University of Leeds, UNITED KINGDOM

Received: July 19, 2015

Accepted: November 2, 2015

Published: November 25, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: NSB was supported by the Russian Science Foundation Grant 14-24-00157. DRU and DNI were supported by ERC grant 335980_EinME. The authors acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208. HCC and PC were supported by the Spanish Ministry of Economy and Competitiveness (including State Training Subprogram: predoctoral fellowships for the training of PhD students (FPI) 2013). MC and

GF were supported by the CRG. The publication cost was covered by CRG.

Competing Interests: The authors have declared that no competing interests exist.

of conformations is astronomical. This fact, known as the Levinthal paradox [10], remained unexplained until the discovery of nucleation mechanism [11]. Nucleation-based model solved the paradox, while predicting that the time required to fold a protein is proportional to $L^{2/3}$, where L is the number of residues [12,13]. In contrast, the influence of the protein topology on the folding was discovered empirically [14]. Developing methods to predict protein folding rate may highlight unknown determinants of protein folding and lead to a detailed understanding of how proteins self-organize.

Predictive methods usually provide an estimate for the natural logarithm of the folding rate of a protein (here referred to as the log folding rate), and they are typically scored using the correlation between the predicted log folding rate and the actual log folding rate. For convenience, we will refer to this score as the "correlation of the model", even though it is a joint property of the model and of the training set. It has long been observed that the folding rate of a protein strongly depends on its length [13,15,16]. Consequently, models that predict the log folding rate using only the length of a protein can reach correlations as high as 0.70 [17]. By using the topology of the protein [14], the correlation can be further improved [18]. Can more sophisticated approaches make better predictions, and if so, what key features must be taken into account?

The ongoing accumulation of experimental data [14,19–22] has accelerated the development of statistical and machine learning methods to address this question [23–34]. Those studies claim correlations ranging from 0.74 [27] to 0.99 [24], among which many lie above 0.90 [24,29–33]{FormattingCitation}. Here, we tested three of those models [24,26,29] against recently collected experimental data. We found much lower predictive powers than the original claims. In all instances, the unifying cause was overfitting, an umbrella term describing situations where models perform well on training data and poorly on new data. Based on this, we suggest that claims of high correlations should be taken with caution and that future studies should demonstrate that they do not suffer from overfitting by using learning curves.

Results

Data set

We collected folding rates obtained experimentally using two references [22,35]. The whole data set contains 113 single-domain proteins without disulphide bonds; 74 of those have two-state folding kinetics in physiological conditions, and the remaining 39 have multi-state kinetics (S1 Table). Here on we refer to this combined data set as "data set 113".

Small sample singularities

In a study by Huang and Tian [26], the authors introduce a parameter Ω for each amino acid, defined as the sum of its rigidity R [36] and its dislike for all regular secondary structures D. D is calculated as a linear combination of parameters P_α, P_β and P_{turn} [37], which measure the resistance of each amino acid type to form α-helix, β-sheet and reverse turn, respectively. The authors estimated the three parameters by fitting a linear regression model on experimental log folding rates. Summing Ω over all the amino acids of the protein, they obtain a total Ω, used as a predictor of the log folding rate. The reported correlation of the model for 40 two-state proteins from 30 to 200 residues long is equal to 0.82 (blue circles in the Fig 1).

We benchmarked the model of Huang and Tian using log folding rates of proteins from data set 113 (red circles in the Fig 1). On this new data set, the model achieved a correlation of 0.63. Here we took care of including only two-state proteins, as the model makes claims for this category only. For the same reason, we included proteins of length comprised between 30 and

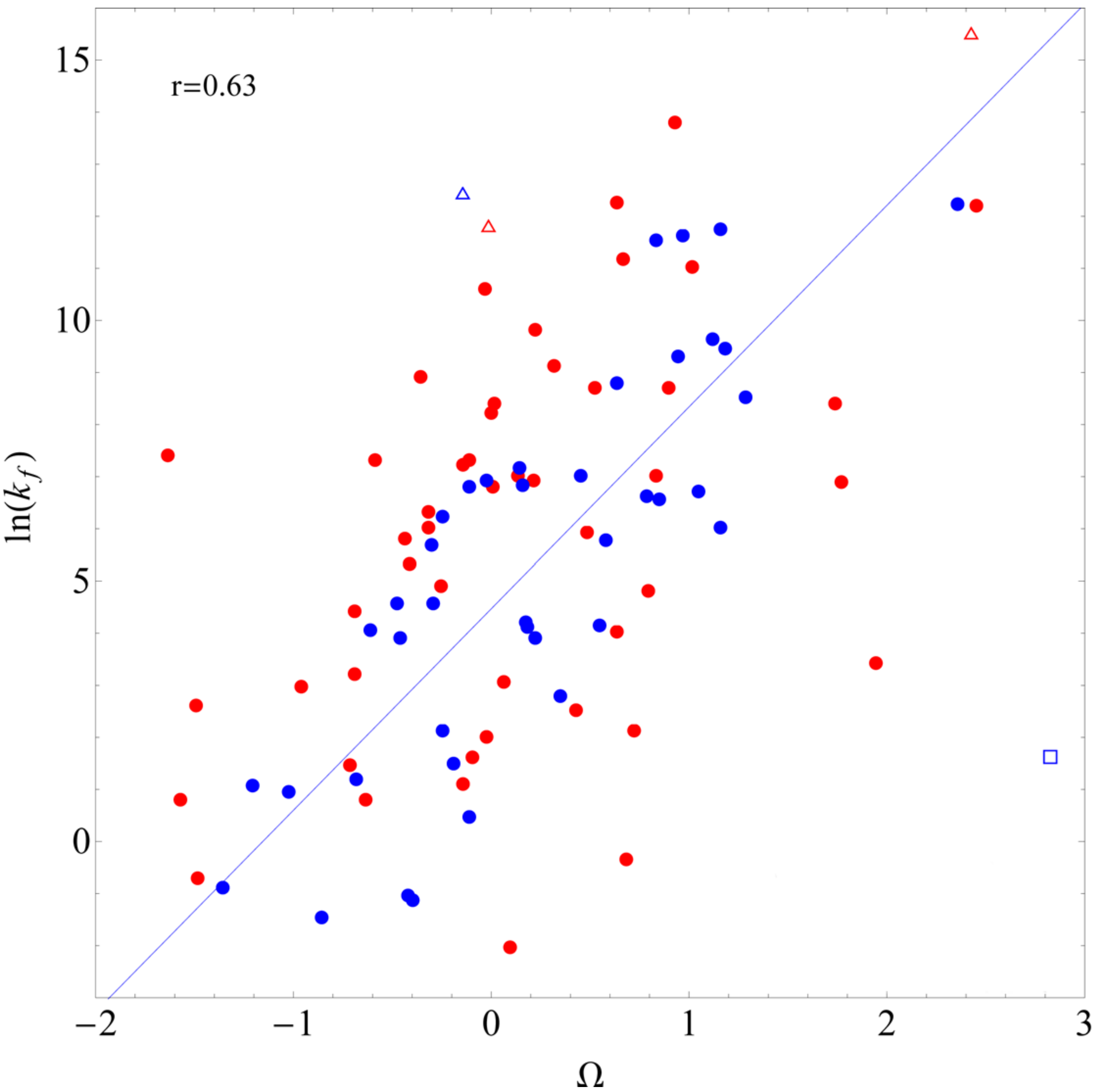

Fig 1. Correlation of Huang and Tian's model. The correlation between Ω, the sum of amino acid foldabilities proposed in [26], and the log folding rates for two-state proteins. Blue dots represent proteins from the data set of Huang and Tian [26]. Red symbols show two-state proteins from data set 113. Correlation coefficients were calculated using only proteins of length comprised between 30 and 200 residues, depicted as circles (0.82 for Huang and Tian's set and 0.63 for two-state proteins from data set 113). Proteins with fewer than 30 amino acid residues are shown as triangles, while those with more than 200 residues are shown as squares. The line shows the prediction from the original model by Huang and Tian [26].

doi:10.1371/journal.pone.0143166.g001

200 residues only. Thus, the discrepancy is not due to extrapolation beyond the domain of validity of the model.

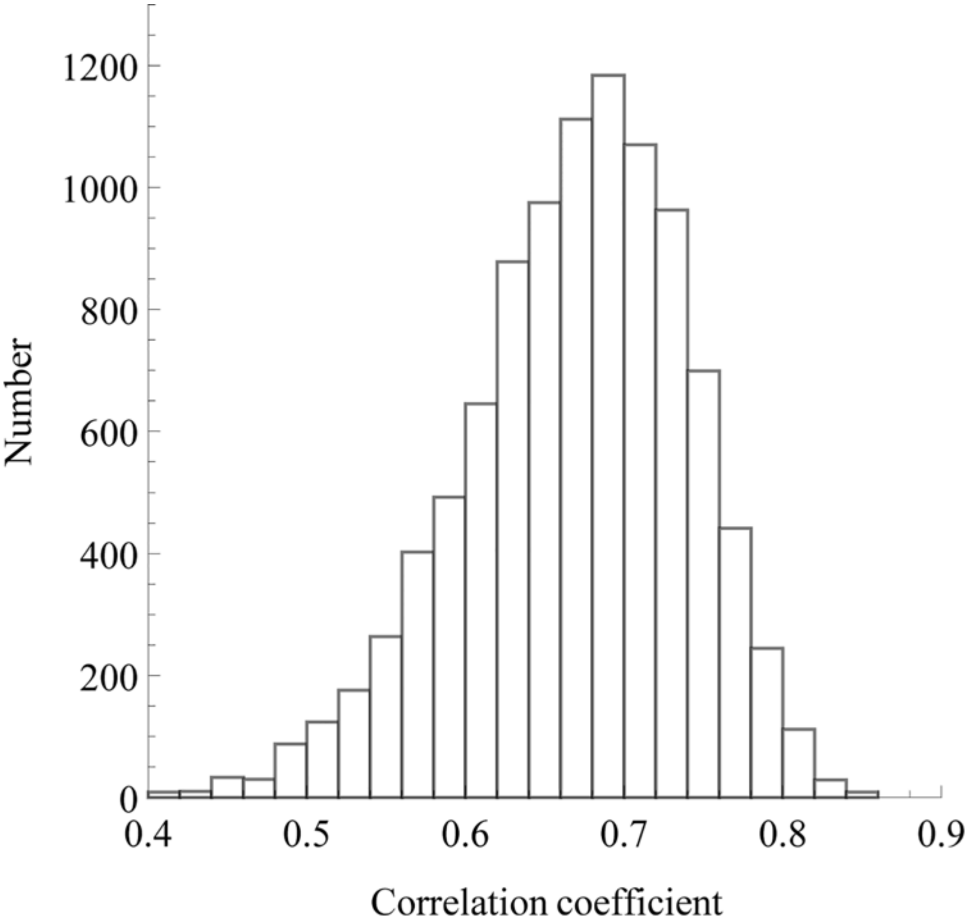

Fig 2. Correlation coefficient of Huang and Tian's model for different samples. Forty data points were randomly sampled from a meta data set and the model described by Huang and Tian [26] was fitted again 10,000 times. The meta data set consists of two-state proteins from 30 to 200 residues combined from [26] and data set 113, without duplicates. The histogram of the obtained correlation coefficients was then plotted. The correlation coefficient ranges from 0.5 to 0.8 approximately, which shows that robust estimation of the correlation cannot be achieved with 40 proteins.

doi:10.1371/journal.pone.0143166.g002

To understand the difference with the original claim, it is important to realize that linear regression is an estimation problem. In other words, the coefficient of correlation is a random variable with an inherent uncertainty due to sampling. A widespread misconception is that unbiased estimators are exact. The fundamental issue here is that the estimators are inaccurate due to the low amount of experimental data in the training set. The correlation can be high because the training set accidently contains an unusually high number of proteins that strictly follow the model of Huang and Tian.

To show this, we combined the original training data set from [26] with data set 113, removed duplicates, and sampled 40 proteins at random (the number of proteins of the original training set) in order to fit again the model of Huang and Tian. As can be seen on Fig 2, the coefficients of correlation thus obtained fluctuate widely; most of the values are comprised in the range 0.5–0.8.

A danger of statistical approaches is that the training sample may not be representative of future data sets. Modelling peculiarities found only in the training set will result in overfitting. With a correlation equal to 0.82, the training set used by Huang and Tian is an outlier

($p = 0.004$), explaining why the predictive power was low on new data. In general, smaller samples have more chances of being aberrant. With 40 data points, the authors may have found a correlation anywhere between 0.5 and 0.8. Unfortunately, one cannot know that a sample used for training is non-representative before acquiring new data. But one can discover that this is *possible* by inspecting the variance of the estimates. Learning curves can be used for this purpose, as we suggest below.

Overtraining

In another study by Gromiha, the author predicts the log folding rates of proteins from physical and conformational properties of their amino-acids [24]. Separating proteins based on their secondary structures ("alpha-helices only", "beta-sheets only" or mixed) he obtains three models, with a range of correlations between 0.95 and 0.99. In a later study by Gromiha, Thangakani and Selvaraj, the authors use the same approach on a larger data set and obtain correlations between 0.90 and 0.99 [29].

We challenged the models of the first study with the proteins of data set 113 and obtained a predictive power much lower than claimed [24]. The coefficients of correlation for categories "alpha-helices only", "beta-sheets only" and mixed were -0.28, 0.008 and 0.02, respectively. The first measure is negative, but not significantly different from 0 (correlation test, p = 0.147), so it should be interpreted as no evidence for statistical association. The last two numbers speak for themselves.

The parameters of the second study are not given explicitly but the authors provide the model as web service [29]. The interface conveniently allows to submit inputs of unknown structural class. Using this option, we submitted the proteins of data set 113 and obtained a correlation coefficient with the measured log folding rates equal to 0.14. In summary, those models have no or very modest predictive power. How to explain that they had spectacular performance on the earlier data sets?

The common issue between those models is overtraining. When the same training set is used many times, the risk is that a model accidentally captures the singularities of the data set. This risk increases when the training set is small and when the number of trainings is large. In the first study for instance, the training set for "beta-sheets only" consisted of 13 proteins [24], and it was used to train over 2 million models (each with 3 to 6 parameters).

Overtraining is a ubiquitous risk in machine learning because it is easy to perform unwittingly. The standard approach to reducing this risk is to separate the data into a training set and a testing set. The training set is used to choose a predictive model, and the testing set is used to evaluate its predictive power. In the variant of this approach called cross-validation, the procedure is repeated and averaged, so that the same data point may be used for training and for testing. Cross-validation is a good statistical practice but not a guarantee, as even cross-validated models may be overtrained.

To illustrate this point, we generated two Gaussian samples with a pseudo-random generator. Being generated separately, the samples are independent. Consistently, the measured correlation between them was -0.0495 (correlation test, $p = 0.625$). When then performed 5-fold cross-validation as in [34] by segmenting the samples in five blocks, using four to train a linear model and measuring the predictive power on the fifth. There are many ways to segment the sample in 5 blocks, and there is as much flexibility to choose a partition where the predictive power is abnormally high. In our example, 1,000,000 tested partitions gave the best correlations equal to 0.20 (positive) and -0.47 (negative), associated p-values without multiple-hypothesis correction equal to 0.044 and $7 \cdot 10^{-7}$, respectively (Fig 3). Claiming that these models have cross-validated correlation equal to 0.20 and -0.47 is true but harmful, since there is in reality

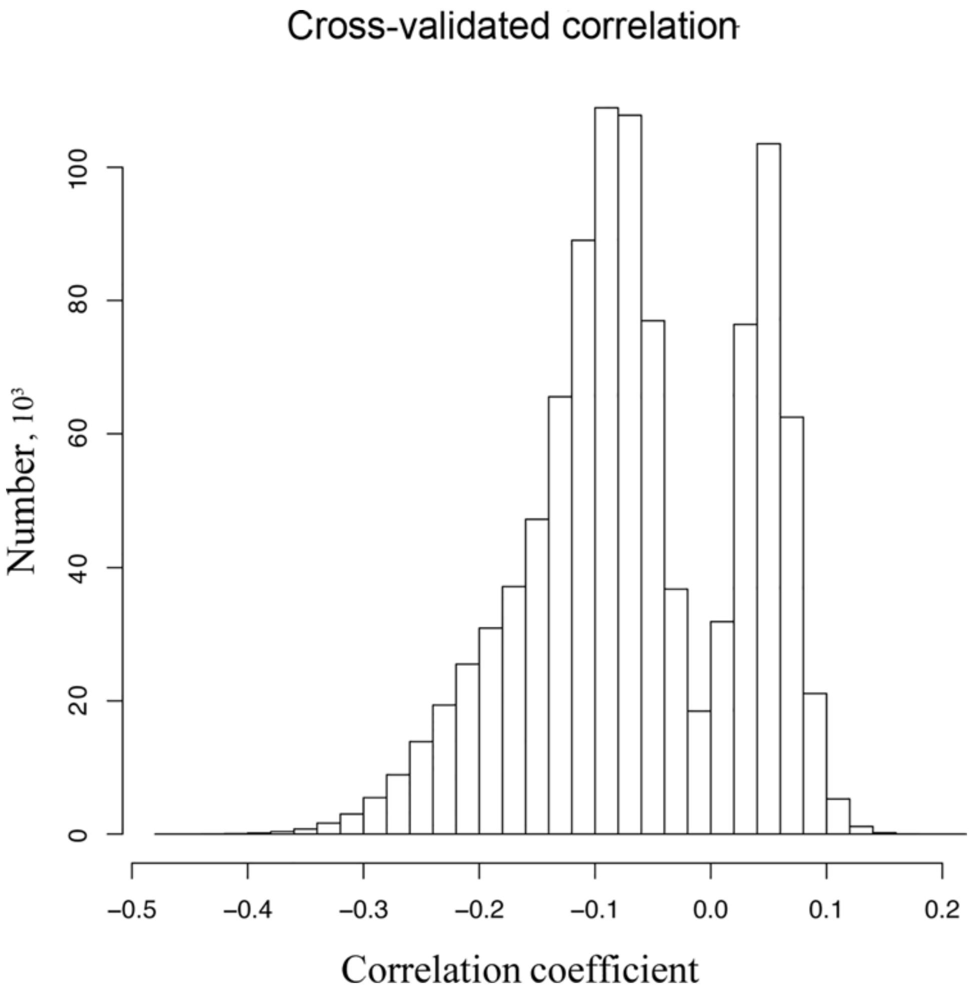

Fig 3. Cross-validation results for two independent Gaussian samples. In this toy model, we try predict a variable from an uncorrelated predictor. The predictive power is null, but the model can be overtrained and give the illusion that the variables are correlated. We repeatedly performed 5-fold cross validation 1,000,000 times on the same data set (n = 100). The plot shows the distribution of the obtained coefficient of correlation. The highest value is 0.202, and the lowest is -0.472 (associated *p*-values without multiple-hypothesis correction equal to 0.044 and $7{\cdot}10^{-7}$, respectively).

doi:10.1371/journal.pone.0143166.g003

no statistical association between the variables. As in other cases of overfitting, the definitive criterion is to measure the predictive power on new data, that is to say data that were never “seen” by the model at any stage of its construction.

How much can be achieved?

How much predictive power can be achieved from statistical and machine learning-based methods is an open question. In particular, it is often impossible to establish a hard limit between the achievable and the non-achievable. In this section, we focus on composition-based predictions of the log folding rate because the number of features is small enough for an exhaustive linear fit. We consider the most complete linear model based on amino-acid composition, that is to say, the one that consists of one parameter per amino-acid. In this model, each amino-acid type brings its own contribution to the log folding rate of the protein. This model contains 21 parameters and extends *every* linear model based on amino-acid composition.

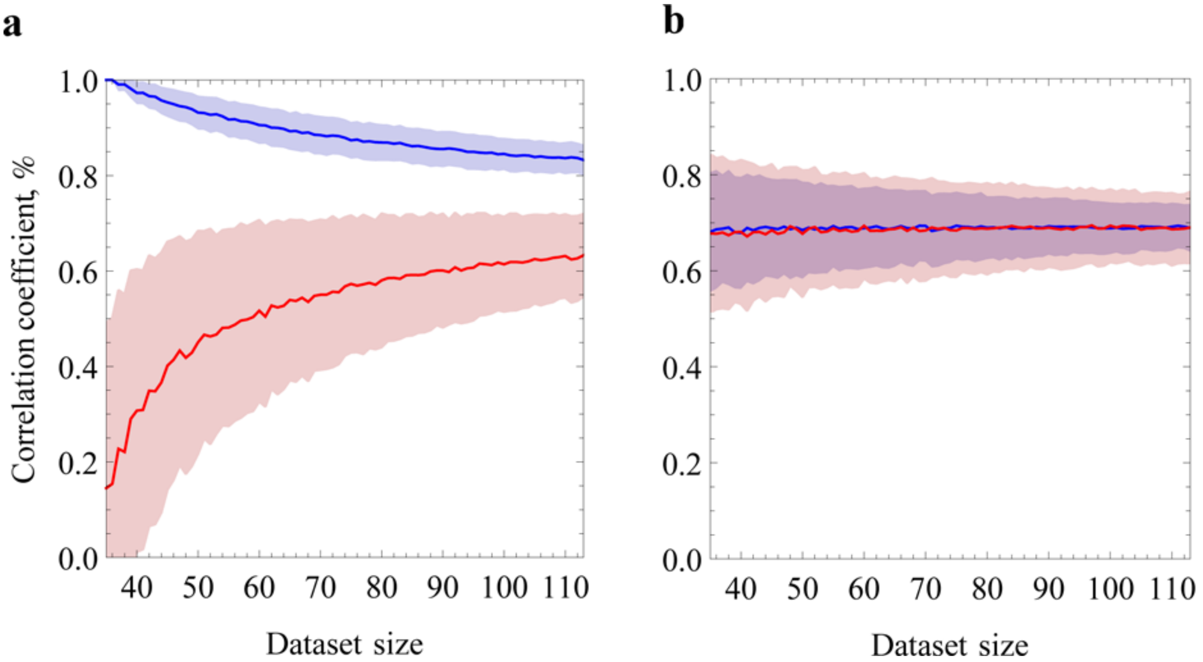

Fig 4. Learning curves of the linear regression model. The mean (n = 1000) correlation coefficient of the training and test sets between the predicted and observed log folding rates (blue and red lines, respectively) is plotted as a function of the dataset size, together with the standard deviations of both sets (blue and red regions, respectively). Sixty percent of the examples are assigned to the training set and 40% to the test set. **a.** Log folding rates were fitted with 20 features corresponding to the absolute amino acid frequency of each protein. A clear overfit can be seen as a gap between the two correlation lines. **b.** Log folding rates were fitted using a single feature corresponding to the amino acid length of each protein to the power of 2/3, $\ln(k_f) \sim -L^{2/3}$ [13]. There exists a nearly-perfect correspondence between training and test sets, and a slightly higher correlation on the test set than in Fig 4A.

doi:10.1371/journal.pone.0143166.g004

We fitted this model with an increasing number of data points and plotted the learning curves in Fig 4A. The learning curve consists of the amount of explained variance R^2 (the square of the coefficient of correlation) on a test data set and on the data set used for training. As the sample size increases, there are more points to fit in the training set with the same number of parameters, so the explained variance decreases (blue). Meanwhile, the model becomes more general and acquires more predictive power on new data, so the explained variance on the testing set increases (red). The vertical distance between the two lines shows the extent of overfitting, or in other words the lack of fit of the model when confronted with new data. When the two curves meet, the model is not overfitted and the true predictive power is the value of the common asymptote. Note that for consistency with the previous sections, we plotted the correlation coefficient instead of the more common R^2. The two lines do not converge, even when all the data points available are included, which means that there is presently not enough experimental data to properly train the complete model.

On the other hand, we used one of the simplest existing model to fit the same data, namely the nucleation-based model $\ln(k_f) \sim -L^{2/3}$ [13], which only takes into account the size of the protein L. For this model, the learning curve shows that the explained variances on the training and test sets are indistinguishable (Fig 4B), which means that the model is not overfitted (for $L^{1/2}$ [16] and $\ln(L)$ [15] the curves are indistinguishable as well, data not shown). However, the performance corresponds to a correlation around 0.70 with a significant uncertainty around this value even when all the data points are included.

Using the same learning curve approach, we tested slightly more complicated models based on contact order [14,18,38]. Absolute contact order is the average number of residues separating by chain the atoms contacting in protein structure [18,38]. Relative contact order is a further normalization of absolute contact order by the number of residues, thus describing the average fraction of protein residues separating the atoms contacting in the structure [14]. For the purpose of calculating contact order, atoms are assumed to be in contact if they are closer than $d = 6Å$ and the chaining distance between corresponding residues $\Delta L \geq 1$, which is reasonable from the physical point of view. With fixed parameters d and ΔL, the learning curves are indistinguishable, as with $L^{2/3}$ (Fig 5A and 5B). If parameters d and ΔL are allowed to vary, the learning curves diverge slightly (Fig 5C and 5D). This slight overfit means that even for a model with three parameters, the currently available amount of experimental data can be an issue.

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 few parameters.

Discussion

The abstraction of protein folding rate has reduced the complex process of protein folding to a single number, making it easy to formalize for machine learning tasks. The regular releases of experimental folding rates [14,19–22,35] makes the prediction of protein folding rates a tempting task, especially with the use of machine learning techniques, where many models can be proposed regardless of their interpretation. The additional possibility to split proteins into different structural classes or into two- and multi-state proteins (which proved initially useful in understanding protein folding principles [14,18]) makes the task even easier. While these tasks are easy to perform, it is equally easy to make a mistake while performing them.

When challenging published models with new data, we discovered that the claims to predict the log folding rate with a correlation higher than 0.90 were too optimistic because the models were overfitted. More precisely, smallish data set and overtraining were the major sources of overfitting. It is worth mentioning that we did not find any example of overparameterization, which is another well-known pitfall of machine learning. With 113 experimental folding rates, the data set used in this study is one of the largest available. This is far from "big data". Methods of prediction and feature extraction that have proved successful for larger data sets may not be directly applied to folding rates. Or, more correctly, not until more data is available. In the current context, hypothesis-driven approaches are more called for.

The established determinants of protein folding are protein size and topology [13–16,18]. We argue that the low amount of experimental data currently prohibits discovering more subtle determinants of protein folding rates by statistical and machine-learning methods. The scarcity of experimental data makes it easier to be the victim of overfitting. 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.

Methods

Data set

We collected folding rates obtained experimentally using two references [22,35]. The whole data set contains 113 single-domain proteins without disulphide bonds ("data set 113"). Seventy four of those have two-state folding kinetics at physiological conditions, and the remaining 39 have multi-state kinetics (S1 Table). In order to reproduce the method developed in [26], we took amino acid compositions, sum of Ω´s and log protein folding rates from the Suplementary Table 2 of ref. [26] for 42 records used therein. Huang and Tian excluded two proteins from the final fit, thus leaving 40 proteins for analysis.

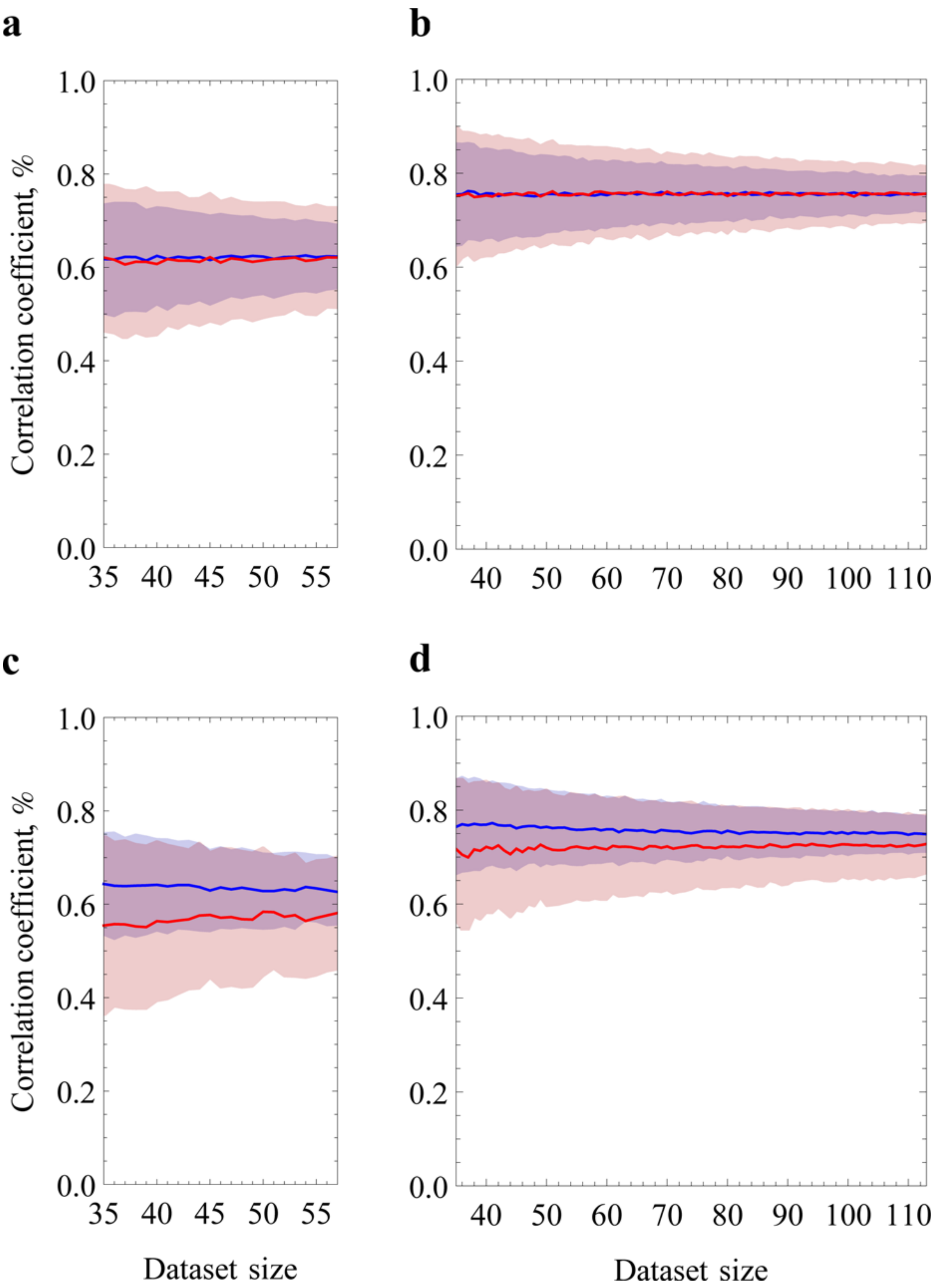
a
b
c
d
Correlation coefficient, %
Correlation coefficient, %
1.0
0.8
0.6
0.4
0.2
0.0
35
40
45
50
55
60
70
80
90
100
110
Dataset size
Dataset size

Fig 5. Learning curves of the contact order models. a. Relative contact order model with fixed parameters d and ΔL (atoms contact in three-dimensional protein structure if they are closer than d = 6Å and belong to the residues having distance by chain $\Delta L \geq 1$). **b**. Absolute contact order model with fixed parameters d and ΔL. Relative (**c**) and absolute (**d**) contact order models with varying parameters d and ΔL. For relative contact order model we restrict the data set to two-state proteins having less than 150 residues.

doi:10.1371/journal.pone.0143166.g005

Analyses

For one-parameter linear regression fit we used "foldability" Ω. Huang and Tian introduced this parameter as a sum of amino acid rigidity and its dislike for all regular secondary structures [26]. They also determined Ω for each amino acid type. Summing Ω values of all residues of protein we calculated the total Ω. FOLD-RATE [29] was queried by a custom Bash script performing an HTTP POST request containing the sequence of the protein of interest and parsing the html response from the server. Briefly, wget was run with options http://psfs.cbrc.jp/cgi-bin/fold-rate/foldrateCalculator.pl—postdata = "sequence = $seq&eqn = unknown". All the linear regression analyses were performed using the lm() function in R with default parameters. Correlations were likewise computed with the cor() function of R.

To give a lower bound on the number of trainings from [24], we used the following passage from the text "As the single property with the highest r-value is not sufficient for accurate prediction I have combined different amino acid properties with a multiple regression fit. The computation has been carried out with the combinations of two to five amino acid properties". There are 2,138,360 different ways to choose two to five amino acid properties among 49, representing as many different models, with 3 to 6 parameters each.

Supporting Information

S1 Table. The list of proteins used in the paper.
(PDF)

Author Contributions

Analyzed the data: MC PC DRU HCC NSB GJF DNI. Wrote the paper: GJF DNI.

References

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. Science. 2012; 338: 1042–1046. doi: 10.1126/science.1219021 PMID: 23180855

2. Sali A, Shakhnovich E, Karplus M. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J Mol Biol. 1994; 235: 1614–1636. PMID: 8107095

3. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round x. Proteins. 2014; 82 Suppl 2: 1–6. doi: 10.1002/prot.24452 PMID: 24344053

4. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012; 149: 1607–1621. doi: 10.1016/j.cell.2012.04.012 PMID: 22579045

5. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci U S A. 2012; 109: 10340–10345. doi: 10.1073/pnas.1207864109 PMID: 22691493

6. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011; 334: 517–520. doi: 10.1126/science.1208351 PMID: 22034434

7. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, et al. Principles for designing ideal protein structures. Nature. 2012; 491: 222–227. doi: 10.1038/nature11600 PMID: 23135467

8. Qiu L, Pabit SA, Roitberg AE, Hagen SJ. Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros. J Am Chem Soc. 2002; 124: 12952–12953. PMID: 12405814

9. Goldberg ME, Semisotnov GV, Friguet B, Kuwajima K, Ptitsyn OB, Sugai S. An early immunoreactive folding intermediate of the tryptophan synthease beta 2 subunit is a “molten globule”. FEBS Lett. 1990; 263: 51–56. PMID: 1691989

10. Levinthal C. How to Fold Graciously. In: DeBrunner JTP, Munck E, editors. How to Fold Graciously. University of Illinois Press. 1969;pp. 22–24.

11. Matouschek A, Kellis JT, Serrano L, Fersht AR. Mapping the transition state and pathway of protein folding by protein engineering. Nature. 1989; 340: 122–126. PMID: 2739734

12. Finkel'shteĭn AV, Badretdinov AI. [Physical reasons for rapid self-organization of a stable spatial protein structure: solution of the Levinthal paradox]. Mol Biol (Mosk). 1997; 31: 469–477.

13. Finkelstein AV, Badretdinov AY. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des. 1997; 2: 115–121. PMID: 9135984

14. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol. 1998; 277: 985–994. PMID: 9545386

15. Gutin A, Abkevich V, Shakhnovich EI. Chain Length Scaling of Protein Folding Time. Phys Rev Lett. 1996; 77: 5433–5436. PMID: 10062802

16. Thirumalai D. From Minimal Models to Real Proteins: Time Scales for Protein Folding Kinetics. J Phys I. 1995; 5: 1457–1467.

17. Ivankov DN, Bogatyreva NS, Lobanov MY, Galzitskaya O V. Coupling between properties of the protein shape and the rate of protein folding. PLoS One. 2009; 4: e6476. doi: 10.1371/journal.pone.0006476 PMID: 19649298

18. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Protein Sci. 2003; 12: 2057–2062. PMID: 12931003

19. Jackson SE. How do small single-domain proteins fold? Fold Des. 1998; 3: R81–R91. PMID: 9710577

20. Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins. 2003; 51: 162–166. PMID: 12660985

21. Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci U S A. 2004; 101: 8942–8944. PMID: 15184682

22. Bogatyreva NS, Osypov AA, Ivankov DN. KineticDB: a database of protein folding kinetics. Nucleic Acids Res. 2009; 37: D342–D346. doi: 10.1093/nar/gkn696 PMID: 18842631

23. Zhang L, Sun T. Folding rate prediction using n-order contact distance for proteins with two- and three-state folding kinetics. Biophys Chem. 2005; 113: 9–16. PMID: 15617806

24. Gromiha MM. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model. 2005; 45: 494–501. PMID: 15807515

25. Ma B-G, Guo J-X, Zhang H-Y. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins. 2006; 65: 362–372. PMID: 16937389

26. Huang J-T, Tian J. Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins. 2006; 63: 551–554. PMID: 16477599

27. Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. Bioinformatics. 2007; 23: 385–386. PMID: 17138584

28. Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem. 2009; 30: 772–783. doi: 10.1002/jcc.21096 PMID: 18752216

29. Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res. 2006; 34: W70–W74. PMID: 16845101

30. Shao H, Peng Y, Zeng Z-H. A simple parameter relating sequences with folding rates of small alpha helical proteins. Protein Pept Lett. 2003; 10: 277–280. PMID: 12871147

31. Zhang L, Li J, Jiang Z, Xia A. Folding rate prediction based on neural network model. Polymer. 2003; 44: 1751–1756.

32. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. J Mol Biol. 2003; 327: 1149–1154. PMID: 12662937

33. Huang L, Gromiha MM. Analysis and prediction of protein folding rates using quadratic response surface models. J Comput Chem. 2008; 29: 1675–1683. doi: 10.1002/jcc.20925 PMID: 18351617

34. Gromiha MM. Multiple contact network is a key determinant to protein folding rates. J Chem Inf Model. 2009; 49: 1130–1135. doi: 10.1021/ci800440x PMID: 19338373

35. Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, Finkelstein AV. Golden triangle for folding rates of globular proteins. Proc Natl Acad Sci U S A. 2013; 110: 147–150. doi: 10.1073/pnas.1210180110 PMID: 23251035

36. Karplus P, Schulz G. Prediction of chain flexibility in proteins. Naturwissenschaften. 1985; 72: 212–213.

37. Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry. 1978; 17: 4277–4285. PMID: 708713

38. Grantcharova V, Alm E, Baker D, Horwich AL. Mechanisms of protein folding. Curr Opin Struct Biol. 2001; 11: 70–82. PMID: 11179895