Briefings in Bioinformatics, 22(1), 2021, 601–603

doi: 10.1093/bib/bbz168
Advance Access Publication Date: 28 December 2019
Letter to Editor

On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation

Castrense Savojardo, Pier Luigi Martelli, Rita Casadio and Piero Fariselli

Corresponding authors: Pier Luigi Martelli, Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, via Belmeloro 6, 40126 Bologna, Italy. Tel.:+39 05120914005; E-mail: pierluigi.martelli@unibo.it; Piero Fariselli, Department of Medical Sciences, University of Torino, via Santena 19, 10126 Torino, Italy. Tel.: +39 0116705071; E-mail: piero.fariselli@unito.it

Abstract

A review, recently published in this journal by Fang (2019), showed that methods trained for the prediction of protein stability changes upon mutation have a very critical bias: they neglect that a protein variation (A- > B) and its reverse (B- > A) must have the opposite value of the free energy difference ($\Delta\Delta G_{AB} = -\Delta\Delta G_{BA}$). In this letter, we complement the Fang's paper presenting a more general view of the problem. In particular, a machine learning-based method, published in 2015 (INPS), addressed the bias issue directly. We include the analysis of the missing method, showing that INPS is nearly insensitive to the addressed problem.

Recently, in this journal, a critical review on machine learning algorithms for the prediction of the protein stability changes upon variations has been published [1]. In the paper, a set of 125 single-site protein variants, with their reverse variations, have been used to evaluate the performance of five methods. The review showed that all algorithms suffer from an overfitting problem [1]. Although the tested methods can suffer from overfitting, the critical point arises because the physics of the system requires that there must be an intrinsic anti-symmetric property of the free energy changes upon variation ($\Delta\Delta G$). Two proteins A and B differing by one residue are each one a variant of the other. Thus, the following relation must hold: $\Delta\Delta G_{AB} = -\Delta\Delta G_{BA}$ (where $\Delta\Delta G_{XY}$ is the free energy change upon single-point variation of protein X that gives rise to protein Y). From this, a predictor of free energy changes upon variations has to fulfill the property: $\Delta\Delta G_{AB} + \Delta\Delta G_{BA} \cong 0$.

This very relevant problem has been introduced for the first time before by the same author [2–4], and it has been recently extended to test the anti-symmetric property ($\Delta\Delta G_{AB} + \Delta\Delta G_{BA} \cong 0$) also when the experimental values are not available [5–7]. This can be obtained by computing the average bias ($<\delta> = \sum(\Delta\Delta G_{dir} + \Delta\Delta G_{inv})/N$) and the linear correlation coefficient (R(Pdir,Pinv)) between the predicted $\Delta\Delta G$ values of the direct and the corresponding inverse variations [5, 6]. Here, we indicate with 'dir' and 'inv' the direct and inverse variations and with the prefix 'E' and 'P' the experimental and the predicted variations, respectively.

A perfect predictor must achieve a correlation between predicted and experimental values close to one (R(Pdir+Pinv, Edir+Einv) = 1), which implies a bias close to zero ($<\delta> = 0$) and a correlation coefficient between direct and inverse variations close to −1 (R(Pdir,Pinv) = −1).

Current datasets are dominated by destabilizing variations, and most of the methods (trained on those datasets) tend to perform better on destabilizing rather than on stabilizing variations, as also shown by Fang in this journal [1]. This finding was also thoroughly analyzed by other authors [5–7] and anticipated by several studies [8–10]. In 2015, we developed a

Castrense Savojardo is affiliated with the Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy.
Pier Luigi Martelli is affiliated with the Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy.
Rita Casadio is affiliated with the Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy.
Piero Fariselli is affiliated with the Department of Medical Sciences, University of Torino, Torino, Italy.
Submitted: 25 October 2019; **Received (in revised form):** 26 November 2019

Table 1. Performance of ΔΔG prediction algorithms for mutations and hypothetical reversed mutations

Method	R(Pdir,Edir)	R(Pinv,Einv)	R(Pinv+Pdir,Einv+Edir)	R(Pdir,Pinv)	<δ>	Inconsistency
mCSM*	0.65	−0.04	0.47	−0.15	−1.66	80.8
DUET*	0.65	−0.15	0.48	−0.11	−1.54	73.6
MuPro*	0.97	−0.02	0.57	0.05	−1.85	75.2
iMutant*	0.94	0.05	0.60	−0.09	−2.10	73.6
STRUM*	0.84	−0.06	0.60	0.06	−1.38	77.6
INPS	0.51	0.51	0.67	−0.99	−0.01	3.2

*= taken from reference [1]. R(X,Y) is the correlation coefficient between X and Y. Pdir = predicted direct variations. Pinv = predicted inverse variations. Edir = experimental direct variations. Einv = experimental inverse variations. Inconsistency = the percentage of direct variations and their inverse pairs predicted with the same sign.

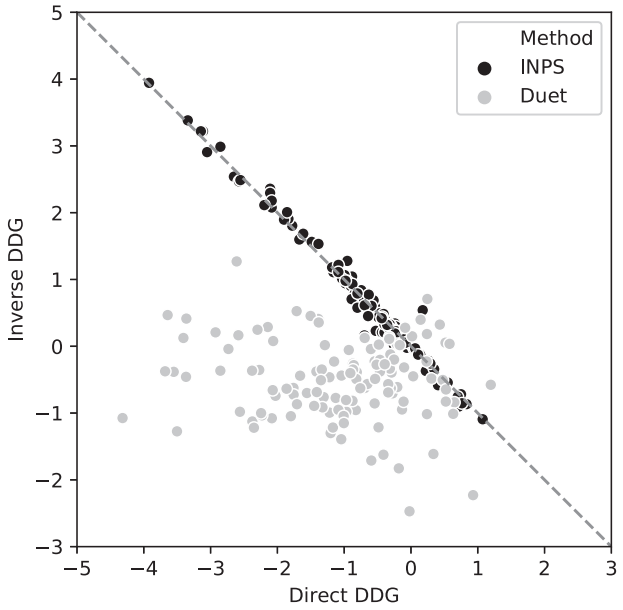

Figure 1. Correlation between the predictions of the direct and inverse variations (using the 125 pairs from [1]). INPS method with respect to the second-best DUET in Table 1. The diagonal line represents the perfect theoretical prediction.

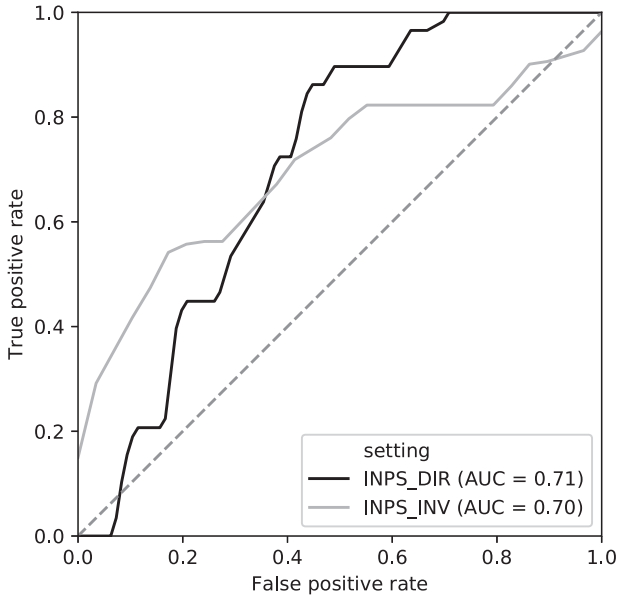

Figure 2. ROC curves for the INPS ΔΔG predictions on the Fangs' dataset [1]. Although INPS was trained to predict real values, here, only the ΔΔG sign is considered. INPS_DIR and INPS_INV are the subsets of direct and inverse variations, respectively.

sequence-based predictor INPS that was designed to take into account the anti-symmetric property [11]. Since INPS was not included in Fang's study published in this journal [1], we report its performance as an addition to the discussion in Table 1 (with the methods evaluated in the original article [1]). It can be seen that the INPS performances are very stable with respect to the direct and inverse mutation, with a very good performance when the average bias (<δ>) and correlation (*R*(Pdir,Pinv)) indices are considered. The meaning of the correlation between direct and inverse predictions (*R*(Pdir,Pinv)) can be explained by the graph reported in Figure 1, where INPS is compared with DUET (the second-best method in Table 1). These results show that a machine learning-based method, properly trained [12], is quite robust with respect to inversion of the variation.

The performance of INPS reported for the direct variations (first column of Table 1) is comparable with those obtained with a proper cross-validation procedure, which does not use similar proteins in training and in testing [5, 7, 11, 12, 13]. However, the variability of the Pearson scores in the range of 0.5–0.7 can also be due to the tested datasets and their distributions [14]. On the contrary, the very high performances achieved by the other methods for the direct variations only (Table 1) indicate skewed training procedures and lack of anti-symmetry. Most of the predictors are thus biased due to the fact that they are not designed to be anti-symmetric by construction, and algorithms are trained using only one part of the information (direct variations).

It is also worth noticing that there is an intrinsic noise in the experimental data: the same variations are characterized by different ΔΔG values (sometimes even with different signs), and this can affect the predictor learning process [14].

Here, we show that when a method is properly trained (using appropriate cross-validations) and when the anti-symmetry is taken into account [5, 12], possibly including input features that are intrinsically anti-symmetric [14, 15], it can produce consistent results for both direct and inverse variations (ROC in Figure 2).

Finally, we would like to add to the final Fang's key point stating that the current methods have not yet reached a perfect maturity and a level for a blind practical use. For us, properly trained methods can provide useful indications when no experimental results are available, as recently confirmed by a small blind test [16]. With this letter, we point out that newly developed machine-learning methods, exploiting the intrinsic anti-symmetric property of the free energy difference and possibly using better and larger datasets, can become useful complements to experimental research.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

P.F. tanks the Italian Ministry for Education, University and Research under the programme "Dipartimenti di Eccellenza 20182022 D15D18000410001", and the PRIN 2017 201744NR8S "Integrative tools for defining the molecular basis of the diseases".

References

1. Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief Bioinform* 2019. doi: 10.1093/bib/bbz071.
2. Li Y, Zhang J, Tai D, *et al*. Prots: a fragment based protein thermo-stability potential. *Proteins* 2012;**80**:81–92.
3. Li Y, Fang J. PROTS-RF: A robust model for predicting mutation-induced protein stability changes. *PLoS One* 2012;**7**:e47247.
4. Fang J. Reliability of machine learning based algorithms for designing protein drugs with enhanced stability. *Drug Designing* 2015;**4**:e130.
5. Pucci F, Bernaerts KV, Kwasigroch JM, *et al*. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* 2018;**34**:3659–65. doi: 10.1093/bioinformatics/bty348.
6. Usmanova DR, Bogatyreva NS, Ariño Bernad J, *et al*. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* 2018;**34**:3653–8. doi:10.1093/bioinformatics/bty340. PMID: 29722803.
7. Montanucci L, Savojardo C, Martelli PL, *et al*. On the biases in predictions of protein stability changes upon variations: the INPS test case. *Bioinformatics* 2019;**35**:2525–7. doi: 10.1093/bioinformatics/bty979.
8. Capriotti E, Fariselli P, Rossi I, *et al*. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* 2008;**9**(Suppl 2):S6.
9. Christensen NJ, Kepp KP. Accurate stabilities of laccase mutants predicted with a modified FoldX protocol. *J Chem Inf Model* 2012;**52**(11):3028–42.
10. Thiltgen G, Goldstein RA. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One* 2012;**7**(10):e46084.
11. Fariselli P, Martelli PL, Savojardo C, *et al*. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* 2015;**31**(17):2816–21.
12. Savojardo C, Fariselli P, Martelli PL, *et al*. INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 2016;**32**:2542–4. doi: 10.1093/bioinformatics/btw192.
13. Montanucci L, Capriotti E, Frank Y, *et al*. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* 2019;**20**(Suppl 14):1–10. doi: 10.1186/s12859-019-2923-1.
14. Montanucci L, Martelli PL, Ben-Tal N, *et al*. A natural upper bound to the accuracy of predicting protein stability changes upon mutations. *Bioinformatics* 2109;**35**(9):1513–7. doi: 10.1093/bioinformatics/bty880.
15. Benevenuta S, Fariselli P. On the upper bounds of the real-valued predictions. *Bioinform Biol Insights* 2019;**23**(13): 1177932219871263.
16. Savojardo C, Petrosino M, Babbi G, *et al*. Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge. *Hum Mutat* 2019;**40**(9):1392–9. doi: 10.1002/humu.23843.