BIOINFORMATICS LETTER TO THE EDITOR Vol. 31 no. 4 2015, pages 614–615
doi:10.1093/bioinformatics/btu682

Structural bioinformatics

Advance Access publication October 17, 2014

Comment on 'protein–protein binding affinity prediction from amino acid sequence'

Iain H. Moal* and Juan Fernández-Recio

Joint BSC-IRB Research Program in Computational Biology, Life Science Department, Barcelona Supercomputing Center, Barcelona, Spain

Associate Editor: Burkhard Rost

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: iainios@hotmail.com

Received and revised on September 7, 2014; accepted on October 14, 2014

1 INTRODUCTION

Predicting the strength of interactions between globular proteins is a central and important topic in structural bioinformatics (Moal *et al*., 2013). The amino acid sequence represents the chemical bonding in a protein which, along with the solvent, dictates how it folds into an ensemble of thermally accessible states. In turn, structure specifies the strength and identity of its binding partners, by establishing the specific arrangements of intermolecular interactions and the intramolecular strain required to achieve them. In a recent paper, Yugandhar and Gromiha (2014) claim to be able to circumvent this and predict protein–protein binding affinities directly from sequence with astounding accuracy. In this letter, we highlight that

(1) Feature selection by stepwise regression is applied without employing an information criterion, y-scrambling or regularization, and the method is not validated using an outer cross-validation loop or external validation set.

(2) The reported energy functions contain many terms that are functionally unrelated to binding.

(3) The reported prediction errors are significantly lower than experimental errors in the training set and variations due to environmental factors.

(4) When an external test set is used, predicted affinities have a correlation of approximately zero with experimental data.

2 STEPWISE REGRESSION IS USED INAPPROPRIATELY

The work employs stepwise regression, a greedy forward selection algorithm. The authors arbitrarily limit the number of parameters to five per model, not including the constant, instead of using the Akaike or Schwarz information criterion or an early stopping set to stop feature selection. Even with these methods, using leave-one-out cross-validation for feature selection can still result in inappropriate components being selected. The algorithm selects from 113 uncorrelated ($r < 0.65$) features. As the training data is split into categories with a median of 12 complexes, the number of parameters outweigh the number of observations by an order of magnitude, making their method highly prone to overfitting. A similar situation was encountered in some of the kinetic rate constant models of Moal and Bates (2012), which also employed stepwise regression. Even when early stopping regularization was used, one of the models, with 27 observations and 200 parameters to select from, gave an inner cross-validation correlation of 1.0 and root mean square error of 0.0. However, in this case an external model selection set was employed and used to reject the model. An alternative validation would be to employ an outer cross-validation loop around the stepwise regression, as in the multivariate adaptive regression spline model reported in Moal *et al*. (2011); the algorithm used an inner bootstrap aggregating loop for feature selection and pruning, with an outer leave-one-out cross-validation loop for validation, supplemented by a final external validation set. Neither of these overfitting avoidance strategies were employed in Yugandhar and Gromiha (2014). Instead, the authors use the same metric for performance evaluation as for selecting features. This gives rise to the discordances below.

3 THE ENERGY FUNCTIONS ARE REPLETE WITH FUNCTIONALLY IRRELEVANT TERMS

The energy functions include many terms calculated using amino acid parameters taken from the AAindex resource (Kawashima *et al*., 2008). These include features that are functionally irrelevant to binding, such as amino acid weights in neural networks for secondary structure prediction, ^{1}H NMR spin-spin coupling constants, and conformational propensities for turns, double bends, helix termini or interdomain linkers. Moreover, the terms are not found consistently between functions; the only term that is selected more than once is the Kerr constant for the amino acids, which is a measure of how refractive index varies in an applied electric field. As none of the selected AAindex features are related to known factors relevant to protein–protein interactions, it is likely that these are selected because they fit the noise.

*To whom correspondence should be addressed.

4 GENERALIZATION ERROR IS LOWER THAN EXPERIMENTAL ERROR IN THE TRAINING SET

In the 72 complexes in the antibody/antigen, non-cognate, G-protein and miscellaneous categories, the cross-validated mean absolute errors (MAE) are reported in the 0.2–0.4 kcal/mol range. This is below the experimental uncertainties estimated in Kastritis *et al.* (2011) and Moal and Fernandez-Recio (2012) by comparing differences in reported affinities determined by different laboratories (around 0.4 kcal/mol). Indeed, for the G-protein and first miscellaneous category, the model error is below the standard deviation typically reported from repeat measurements within the same article using identical solutions, conditions and equipment (up to 0.25 kcal/mol). Similar differences in binding can be found by varying the temperature in a 15 °C ambient temperature range or even changing the buffering agent. Much greater differences arise from changing ionic strength or pH, the latter of which can alter affinity by 1–2 kcal/mol over the 5.5–8.5 range (Kastritis *et al.*, 2011), which is greater than the MAE reported in all functional categories. Further, the interactions used include complexes of sub-nanomolar affinity, as determined using isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR). Both SPR and ITC are prone to error when evaluating interactions of such high strength, not to mention effects due to SPR tethering. Taken together, this indicates that the reported cross-validated prediction correlations (0.74–0.99, mean 0.91) and errors (0.18–1.17 kcal/mol, mean 0.5 kcal/mol), cannot be used to estimate generalization error and are a result of severe overfitting.

5 THE METHOD FAILS ON AN EXTERNAL VALIDATION SET

Even when implemented correctly, high cross-validation performance is a necessary but not sufficient condition for a predictive model (Golbraikh and Tropsha, 2002). The most stringent test of a model is its performance on an external validation set, which provides an objective estimation of predictive value. Such a dataset is shown in the Supplementary Data, derived from Chen *et al.* (2013) by removing interactions involving chains of fewer than 50 residues,

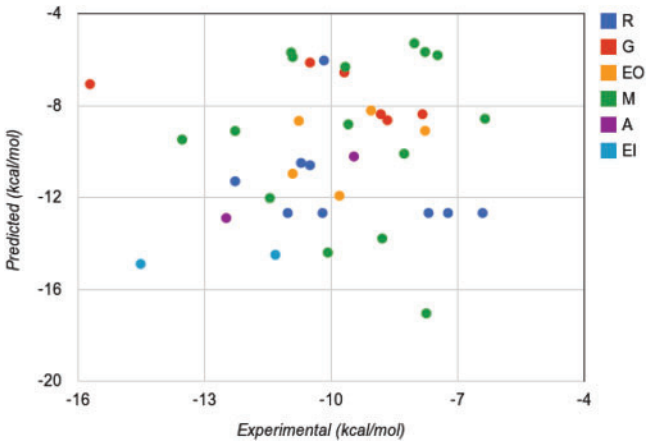

Fig. 1. Predicted versus experimental binding affinities, categorized as receptor (R), G-protein (G), enzyme/inhibitor (EI), enzyme/other (EO), antibody/antigen (A) and miscellaneous (M)

Table 1. Summary of performance on external test set

Category[a]	*N*	*r*	*P*	MAE	r_{BSA}	p_{BSA}
M	15	0.02	0.95	3.43	–0.64	**0.01**
R	9	–0.35	0.37	2.92	–0.14	0.72
G	6	–0.47	0.34	2.86	–0.53	0.29
EO	5	0.34	—	1.29	–0.85	—
A	2	–	—	0.58	—	—
EI	2	–	—	1.78	—	—
All	39	0.07	0.67	2.72	–0.34	**0.03**

[a]Categories as per Figure 1.

as well as complexes which overlap with the training data. These were submitted to the PPA-Pred web server (http://www.iitm.ac.in/bioinfo/PPA_Pred/). For seven complexes, the server failed to return a prediction (2OMZ, 2QNA, 3BEG, 3BLH, 3KNB, 3MCA and 3OIQ). The results for the remaining interactions are shown in Figure 1 and summarized in Table 1. A statistically insignificant overall correlation of $r = 0.07$ ($P = 0.67$, $n = 39$) is observed, with MAE of 2.7 kcal/mol. When looking at individual categories with more than five members, statistically insignificant anti-correlations are observed, with MAE ranging from 2.9 to 3.4 kcal/mol. By contrast, the known weak anticorrelation with BSA, buried surface area (Chen *et al.*, 2013; Kastritis *et al.*, 2011), is observed consistently in all categories, and is statistically significant overall and in the miscellaneous category. In conclusion, the method reported in Yugandhar and Gromiha (2014) has been evaluated on a blind test set and found to have large errors and a correlation of approximately zero, which more accurately reflects predictive value than the initial flawed validation.

Funding: IHM received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement PIEF-GA-2012-327899.

Conflict of interest: none declared.

REFERENCES

Chen,J. *et al.* (2013) Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.*, **22**, 510–515.

Golbraikh,A. and Tropsha,A. (2002) Beware of q2! *J. Mol. Graph. Model.*, **20**, 269–276.

Kastritis,P.L. *et al.* (2011) A structure-based benchmark for protein-protein binding affinity. *Protein Sci.*, **20**, 482–491.

Kawashima,S. *et al.* (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**(Database issue), D202–205.

Moal,I.H. and Bates,P.A. (2012) Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS Comput. Biol.*, **8**, e1002351.

Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

Moal,I.H. *et al.* (2011) Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, **27**, 3002–3009.

Moal,I.H., Moretti,R., Baker,D. and Fernandez-Recio,J. (2013) Scoring functions for protein-protein interactions. *Curr. Opin. Struct. Biol.*, **23**, 862–867.

Yugandhar,K. and Gromiha,M.M. (2014) Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics*, **30**, 3583–3589.