#56 - Binding affinity prediction using a nonparametric regression model based on physicochemical and structural descriptors of the nano-environment for protein-ligand interactions

Luiz Borro[1,2], *Inacio Yano*[2], *Ivan Mazoni*[2], *Goran Neshich*[2]
[1]*University of Campinas* [2]*Embrapa Agriculture Informatics*

ABSTRACT
We propose a new empirical scoring function for binding affinity prediction modeled based on physicochemical and structural descriptors that characterize the nano-environment that encompass both ligand and binding pocket residues. Our hypothesis is that a more detailed characterization of protein-ligand complexes in terms of describing nano-environment as precisely as possible can lead to improvements in binding affinity prediction. Similar hypothesis has already been proven valid in case of nano-environments for protein-protein interfaces[1] and catalytic site residues (yet to be published).

INTRODUCTION
In structure-based virtual screening campaigns, *in silico* protein-ligand complexes are evaluated and ranked according to their estimated binding affinities. Normally the ranking step is performed by using scoring functions, i.e. mathematical models that assess the strength of interaction between two binding partners. However, scoring functions are generally weak predictors of binding affinity mostly because they fail to model properly polar aspects of the protein-ligand interaction[2]. In order to improve binding affinity prediction, we propose an empiric nonparametric predictive model derived from physicochemical and structural descriptors that characterize the nano-environment that encompass both ligand atoms and binding pocket residues.

METHODS
Datasets. In order to ensure an unbiased performance comparison with other related approaches, we used the PDBbind v2007 refined set, which comprises of 1300 diverse protein-ligand complexes with high quality structural and binding data. The refined set was split into two disjoint sets: a training set of 1105 used for fitting the predictive models; and a test set of 195 complexes (known as core set) for performance evaluation.
Protein-Ligand complex characterization. A given protein-ligand complex is represented by physicochemical and structural parameters from the nano-environment covering the ligand atoms and binding pocket residues. In order to obtain a more detailed characterization, special attention was given to descriptors related to the hydrophobic effect as well as to polar aspects of the protein-ligand binding. Descriptors were divided into three classes: Ligand-Only (7 descriptors), Protein-Only (6 descriptors) and Protein-Ligand (9 descriptors), as shown in Table 1. Protein-Only descriptors and Protein-Ligand descriptors were calculated through the STING platform[3], whereas the Ligand-Only parameters were calculated using Biovia Pipeline Pilot.

Table 1. *List of descriptors used to characterize protein-ligand complexes.*

Class	Descriptors
Ligand-Only	Volume, Polar Solvent-Accessible Surface Area, Strain Energy, Number of Hydrogen Bond (HB) donors, Number of HB Acceptors, AlogP, Number of Rotatable Bonds
Protein-Only	Hydrophobicity, Electrostatic Potential @ Surface, Unused Contacts Energy (HB, Charged, Hydrophobic, Aromatic)
Protein-Ligand	Protein-Ligand Interaction (HB, Charged, Hydrophobic, Aromatic), Ligand Buried Surface, Energy Density, Sponge, Density, Protein Hydrophobicity Variation

Binding affinity prediction model. Using the descriptors listed on Table 1 and the experimental pKi of the training set complexes as input data, the binding affinity predictive model (herein called STING^{SF}) was trained as a regression-based random forest.

RESULTS & CONCLUSIONS
STING^{SF}'s performance was evaluated on the PDBbind benchmark v2007. Table 2 presents a performance comparison between STING^{SF} and the top four previously tested scoring functions on the same benchmark. Clearly our predictive model ranks among the best with regard to binding affinity correlation, having a slightly inferior result in terms of R_P when compared to RF-Score::Elem-v2.
By statistically analyzing the contribution of each descriptor in the predictive model, we observed that the most important descriptors are related to shape complementarity (Ligand Buried Surface Area), hydrophobic effect (Hydrophobicity, ALogP) and polarity (Polar Solvent-Accessible Surface Area, Electrostatic Potential @ Surface). That result may suggest that STING^{SF} can be further improved by expanding the characterization of protein-ligand complexes in terms of hydrophobicity and polarity complementary descriptors. Finally, considering STING^{SF}'s performance on the PDBbind benchmark v2007, the *de facto* standard for validation of scoring functions, we believe that our binding affinity predictive model can be a viable option for rescoring purposes in virtual screening campaigns.

Table 2. *Performance of scoring functions on the PDBBind benchmark as measured by Pearson's correlation coefficient (RP), Spearman's correlation coefficient (RS) and the standard deviation of the difference between predicted and measured binding affinity (SD). Data extracted from Reference 4*

Scoring Function	R_P	R_S	SD
RF-Score::Elem-v2	0.803	0.797	1.54
STING^{SF}	**0.798**	**0.798**	**1.54**
SFCscore^{RF}	0.779	0.788	1.56
RF-Score::Elem-v1	0.776	0.762	1.58
X-Score::HMScore	0.644	0.705	1.83

ACKNOWLEDGEMENTS
The authors thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support (Grant #2015/00428-6).

REFERENCES
1. de Moraes, F. R. et al. *PloS one* 9, e87107 (2014).
2. Li, Y. et al. *Journal of Chemical Information and Modeling 54*, 1717-1736 (2014)
3. Neshich, G. et al. *Nucleic Acids Research* 33, W29-W35, (2005).
4. Ballester, P. J. et al. *Journal of Chemical Information and Modeling* 54, 944-955 (2014)

#57 - Who is my neighbor ? – Continuously evaluating residue-residue contact predictions

Juergen Haas[1,2], Tobias Thuering[1,2], Dario Behringer[1,2], Torsten Schwede[1,2]
[1]SIB Swiss Institute of Bioinformatics [2]Biozentrum, University of Basel

Protein structure modeling is widely used in the life science community to build models for proteins, where no experimental structures are available. The Continuous Automated Model EvaluatiOn platform (CAMEO, http://www.cameo3d.org) currently assesses the performance of servers predicting protein structures (3D) and servers estimating local model quality (QE). (1)

Continuous assessment of e.g. structure prediction servers allows to retrospectively analyze their performance, as the quality of models may vary significantly among different modeling servers depending on the specific target protein and the applied modeling approach. Here, we introduce a new category "Contact Prediction" (CP) to CAMEO assessing residue-residue contact predictions. It was recently shown that the quality and hence utility of a model can be improved greatly by considering residue-residue contact predictions in the modeling process. (2) This applies in particular for target proteins larger than 250 amino acid residues, with no templates available, where commonly comparative approaches fail to produce any model and *de-novo* approaches struggled to produce a meaningful prediction.

CAMEO firstly supports the developers of prediction servers, rapidly assessing new developments anonymously and monitoring the performance of their public productive servers continuously. Secondly, CAMEO also stimulates the respective communities in discussing new scores, thereby covering yet another aspect of the respective field.

CAMEO is based on the PDB weekly pre-release of experimental protein structures, where on Saturday sequences are sent to the participating servers for CAMEO 3D and CP categories. The QE category relies on 3D coordinates which are either harvested from public modeling servers or selected from decoy sets. Four days later the structures are released by the PDB and used by CAMEO as reference for scoring the predictions. Shortly after the assessments are then published on cameo3d.org.

Making evaluation processes available to other communities. The workflow of CAMEO is currently being abstracted in the context of ELIXIR EXCELERATE framework to apply the concepts of this successful evaluation platform to other communities, assessing tasks such as text-mining or multiple alignments of protein sequences.

References
1. Haas, J. et.al. Database **2013**, 10.1093/database/bat031
2. Monastyrskyy B. et.al. Proteins **2016**, 10.1002/prot.24943