Journal of **Chem**informatics

Combining SFCscore with Random Forests leads to improved affinity prediction for protein-ligand complexes

D Zilian, CA Sotriffer[*]

From 8th German Conference on Chemoinformatics: 26 CIC-Workshop
Goslar, Germany. 11-13 November 2012

SFCscore is a collection of emprirical scoring functions derived from a set of over 60 descriptors for protein-ligand complexes of known structure [1]. By the time of their derivation, SFCscore functions were the best-performing scoring functions tested on large heterogeneous data sets, but the overall correlation was still not within the desired range. Similarly, despite the ever increasing amount of structure and affinity data, the general advancements in the development of empirical scoring functions have been rather moderate over the past years. However, more recently, Ballester and Mitchell [2] published a function that outperformed current state-of-the-art scoring functions when tested against the PDBbind benchmark set [3]. This function uses relatively simple atom contact counts as descriptors and is derived by the Random Forest algorithm. Here, we present a study in which we used Random Forests to derive a new function ("SFCscoreRF") based on the SFCscore descriptors as input data. Although this is not a fully non-parametric approach, the descriptors are supposed to capture more accurately the physically relevant interactions. We tested the new function against the PDBbind benchmark set and the CSAR-NRC HiQ 2010 set [4] and, in addition, performed the Leave-Cluster-Out validation as proposed by Kramer and Gedeck for the PDBbind set [5]. The results suggest that the new function significantly improves the predictive power of SFCscore, as it increases the correlation between predicted and experimentally determined affinities for the PDBbind benchmark set from $r^2 = 0.41$ (best previous SFCscore function) to $r^2 = 0.61$ (SFCscoreRF) and for the CSAR data set from $r^2 = 0.38$ to $r^2 = 0.53$.

References

1. Sotriffer CA, Sanschagrin P, Matter H, Klebe G: **SFCscore: Scoring Functions for Affinity Prediction of Protein-ligand Complexes'.** *Proteins: Struct Funct Bioinf* 2008, **73**:395-419.
2. Ballester PJ, Mitchell JBO: **A Machine Learning Approach to Predicting Protein-ligand Binding Affinity with Applications to Molecular Docking.** *Bioinformatics* 2010, **26**:1169-1175.
3. Cheng T, Li X, Li Y, Liu Z, Wang R: **Comparative Assessment of Scoring Functions on a Diverse Test Set.** *J Chem Inf Model* 2009, **49**:1079-1093.
4. Dunbar JB, Smith RD, Yang C-Y, Man-Un Ung P, Lexa KW, Khazanov N, *et al*: **CSAR Benchmark Exercise of 2010: Selection of the Protein-ligand Complexes.** *J Chem Inf Model* 2011, **51**:2036-2046.
5. Kramer C, Gedeck P: **Leave-cluster-out Cross-validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets.** *J Chem Inf Model* 2010, **50**:1961-1969.

* Correspondence: sotriffer@uni-wuerzburg.de
Institute of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, 97074 Wuerzburg, Germany