A Diverse Benchmark Data Set for the Validation of Scoring Functions based on 3D Matched Molecular Pairs

Lena Kalinowsky[1], Julia Weber[1], Shantheya Balasupramaniam[2], Knut Baumann[2], Ewgenij Proschak[1]

1. *Institute of Pharmaceutical Chemistry, Goethe University Frankfurt, Max-von-Laue Str. 9, Frankfurt a.M., D-60438 Germany.*
2. *Institute of Medicinal and Pharmaceutical Chemistry, Technical University of Braunschweig, Beethovenstr. 55, Braunschweig, D-38106 Germany.*

Prediction of protein-ligand interactions and their corresponding binding free energy is a challenging task in structure-based drug design and related applications. Docking and scoring is a broadly used approximation of the latter. To demonstrate the predictive power and to investigate the strength and weaknesses of scoring functions several benchmark test sets have been developed in the past.[1]-[3] These data sets are characterized by high diversity in terms of protein families, ligand chemotypes and binding affinities. High diversity is well suited for the evaluation and comparison of the global performance of docking and scoring software. However, understanding the local behavior of a scoring function, how well it can differentiate between similar molecules is almost impossible with these data sets. Here, a novel benchmark data set based on Matched Molecular Pairs (MMPs) was developed to study the local behavior of scoring functions. MMPs are defined as molecules that differ in one well-defined transformation that is associated with a change in an arbitrary molecular property (transformation effect).[4] The assembled data set of 99 3D-MMP was used to investigate whether or not scoring functions can differentiate between chemically related compounds. Various scoring functions were used to score the data set, most of them are available within the commercially available software MOE 2014.09 [5] and GOLD Suite 5.2.2 [6]. The 3D-MMPs were scored in the respective crystal structures without any posing (i.e. the position of the small molecule was not changed) to focus on scoring and to exclude the influence of posing (i.e. the placement algorithm). Only three scoring functions (X-Score, Affinity dG and GoldScore) reached a prediction rate of more than 60% in the prediction of the trend of a transformation effect. Analyzing the relationship between molecular size and affinity led to the following results. In 51 3D-MMPs, the larger molecule was also the more active one. In only 20 3D-MMPs the smaller molecule was more active (the remaining molecules show no difference in bioactivity (n = 5) or the same number of heavy atoms (n = 23)). Hence, in 71.8% (51 out of 71) the larger molecule was also more active. This means that in this study the molecule's size would be the best scoring function.

[1] B. Kramer, *Proteins Struct. Funct. Genet.*, **1999**, *37*, 228–241.
[2] J. W. M. Nissink, *Proteins Struct. Funct. Genet.*, **2002**, *49*, 457–471.
[3] R. Wang, *J. Med. Chem.*, **2003**, *46*, 2287–2303.
[4] A. G. Leach, *J. Med. Chem.*, **2006**, *49*, 6672–6682.
[5] Chemical Computing Group Inc., Molecular Operating Environment (MOE) 2014.09, **2014**.
[6] CCDC Gold Suite 5.2.2. Cambridge, United Kingdom.