Machine learning models for guiding protein structure selection lead to a boost in the performance of ensemble docking

Ningning Fan,[1] Conrad Stork,[1] Sarah Gritzka,[1] Christina de Bruyn Kops,[1] Johannes Kirchmair[1]*

[1]Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Hamburg, 20146, Germany

The use of multiple structures of target proteins for ligand docking ("ensemble docking") is an established concept in structure-based virtual screening and advantageous in particular when addressing malleable target proteins.[1] In most ensemble docking approaches, a ligand is docked against multiple receptor conformations and the highest-ranked docking pose obtained with any of these protein structures is reported as the result.[2] However, the performance of ensemble docking can often be further improved by manually selecting the most relevant protein structures for screening.[3]

In this contribution, using VEGFR-2 as an example, we introduce a new integrated approach that employs machine learning models for identifying the most suitable protein structure for docking each individual compound. We started by collecting 38 relevant protein structures of VEGFR-2 from the PDB and docking 2320 actives and 24,950 decoys from the DUD-E[4] to each of these structures with Glide.[5] The docking performance on the individual protein structures was moderate, with the maximum ROC AUC (Receiver Operating Characteristic Area Under the Curve) value of 0.78.

Based on the docking results for each of the 38 VEGFR-2 PDB structures, we labeled actives and decoys as "correctly scored" or "incorrectly scored" according to the docking score (GlideScore).[5] The labeled compounds, represented by Morgan2 fingerprints, served as input for training individual random forest classifiers for each of the protein structures. The integrated virtual screening approach runs query molecules against all of these classifiers to identify the most suitable protein structure for docking. In contrast to common ensemble docking approaches, compounds of interest are only docked to the protein structure that obtained the highest score out of all of the classifiers. Benchmarking with an independent test set showed that the integrated ensemble docking approach obtained significantly better ROC AUCs and, in particular, higher early enrichment than the stand-alone ensemble docking approach.

[1] R. E. Amaro, W. W. Li, *Curr. Top. Med. Chem.*, **2010**, *10*, 3–13.
[2] S. Rao, P. C. Sanschagrin, J. R. Greenwood, M. P. Repasky, W. Sherman, R. Farid, *J. Comput. Aided Mol. Des.*, **2008**, *22*, 621–627.
[3] M. Rueda, G. Bottegoni, R. Abagyan, *J. Chem. Inf. Model.*, **2010**, *50*, 186–193.
[4] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, *J. Med. Chem.*, **2012**, *55*, 6582–6594.
[5] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, J. L. Banks, *J. Med. Chem.*, **2004**, *47*, 1750–1759.