A machine-learning scoring function for protein-ligand molecular docking

Oscar Emilio Arrúa Arce, Andrej Aderhold, Adriano Velasque Werhli, Karina dos Santos Machado

Universidade Federal do Rio Grande - FURG, UNIVERSIDADE FEDERAL DO RIO GRANDE

Abstract

In the field of drug design, scoring functions are useful for predicting the binding affinity of protein-ligand complexes. Machine learning approaches are showing promising performance as a result of the increasing amount of data regarding biochemical and biophysical processes, obtained from previous experiments. In this work we propose a machine learning based scoring function for protein-ligand molecular docking. This scoring function was developed according to related works, where: from protein-ligand complexes (training set) were obtained features of proteins, ligands and interactions that are considered as attributes; machine learning methods are to use to train models, including feature selection techniques and hyperparameters optimization; and test sets that are used to evaluate the proposed scoring functions models. As training set, we combine the PDBbind 2016 refined and general sets, CSAR-NRC HiQ and Decoys CSAR-NRC HiQ. As attributes we considered AutoDock Vina score and geometrical, SFCscore, solvent-accessible surface area, DeltaVinaRF20, protein primary and secondary structure and Vina features. We also considered specific software to generate features as PaDEL Descriptor, NNScore 2.0 and RDKit. Random Forest and Gaussian Process were compared as machine learning methods, in addition to LASSO to calculate the weights of the attribute's importance and GridSearchCV as a technique to hyperparameters optimization. Thus, the proposed scoring function was evaluated using the CASF-2016 benchmark, based on Scoring, Ranking, Docking and Screening Power. As a result, for CASF-2016 evaluation, the proposed scoring function achieved good results, comparable to the best scoring functions. As Scoring Power, we obtained 0.81 that corresponds to the Pearson correlation coefficient between predicted affinities and experimental measured affinities. For Ranking Power, the proposed scoring function achieves a Spearman correlation coefficient of 0.66 between the ranks based on the predicted affinities values and the experimentally ones. For the Docking Power, the proposed scoring function obtained 86% success rate in identifying the top best-scored ligand binding pose below 2 root-mean-square deviation from the native pose (and 83.8% without native poses). Finally, for Forward Screening Power, the proposed scoring function has a got 26.5% success rate to identifying potential small-molecule ligands for a chosen target protein at the top 1% level (better than all the scoring functions compared in CASF-2016) while for Reverse Screening Power achieve a 18.5% success rate in identifying potential target proteins for a bioactive small-molecule compound at the top 1% level.

Funding: CAPES, CNPq

Link to Video: