POSTER - PROTEINS AND PROTEOMICS

DEEP LEARNING SCORING FUNCTION FOR PROTEIN-LIGAND MOLECULAR DOCKING

Fabiano Mendonça (mendonca.biano@gmail.com)

Yago Mendes Paes (yagomep@gmail.com)

Andrej Aderhold (andrej.aderhold@protonmail.com)

Adriano Werhli (werhli@gmail.com)

Karina Machado (karinaecomp@gmail.com)

Protein-Ligand Molecular Docking is a method that employs computational techniques to predict the best orientation and conformation of a ligand in the binding site of a target receptor. This can be applied for virtual screening to identify which ligands have the potential to bind to the target molecule. Being an in-silico approach, molecular docking reduces the cost of projects searching for drug candidates and accelerates the process of identifying a required ligand to achieve the desired effect. Usually, this process occurs in two stages: initially, the search for the best position/conformation of a ligand into the receptor's binding site, followed by the calculation of estimated binding affinity using a scoring function (SF). Developing an accurate score function is essential for the success of molecular docking studies, as it directly influences the reliability of the predictions made by the docking software. Therefore, score functions are continuously refined and validated using experimental data and benchmark datasets. The SFs used in molecular docking can be classified into different categories: physic-based, knowledge-based, empirical and machine learning.

Between the types of SFs, ML-based SFs have been demonstrating robustness and can be adapted well to specific simulation conditions. These SFs are trained considering different strategys to choose the features and as target attribute experimental values of binding affinities, obtained from databases as PDBBind and DUD-E. Different ML algorithms can be used for training SFs, for example Random Forest (Delta-Vina), Convolutional Neural Networks (DeepDock, GNINA 1.0), among others. There are many studies aiming to improve the accuracy of SFs through the use of deep learning The application of neural networks has shown high potential, significantly reducing the time required for SF calculation, thus allowing testing of a larger number of ligands. In this work, we proposed the implementation of a scoring function using deep learning techniques, using TensorFlow 2 and Keras, with a convolutional neural network architecture. The same network architecture was used for training with two different datasets PDBbind versions v2016 and v2019. PDBbind is a database of protein-ligand complexes that contains various complexes organized by similarity, with experimental binding affinities. Additionally, PDBbind has a benchmark called CASF (Comparative Assessment of Scoring Functions) to evaluate scoring functions. To represent the molecules for training the network, four-dimensional vectors were used. The first three dimensions represent the physical positions of the atoms in the molecules, while the fourth dimension represents different features of the atoms. These features include molecular descriptors such as type of atoms, the number of bonds, and so on. Features are calculated using Openbabel and the charges have been added using UCSF Chimera. During training, each protein-ligand complex is rotated 24 times to make the network more robust, reducing sensitivity to the initial position of the protein-ligand complex. The network evaluation under the scoring power metric of CASF-2016, when comparing the proposed SF with others SFs, all proposed models presented very satisfactory results, being inferior only to DeltaVinaRF20.

Palavras-chave: protein-ligand molecular docking; virtual screening; convolutional neural network; scoring function.