MeetDockOne: team 1 scoring method for protein-protein docking.

François Gravey, Guillaume Delevoye, Paula Milán Rodríguez, Ilyes Abdelhamid, Maxime Borry

Strategy description:

Our objective in this project is to develop a Scoring program, for protein docking poses, that estimates different types of informations about the physical-chemical characteristics of each pose: contact propensities, shape complementarity, electrostatic energy and Lennard-Jones potential.

Once they are calculated, each set of data is evaluated by a machine learning algorithm that assigns a TM score to each pose. In this way, we can select the best sampling proposals for each complex.

When the program is launched, it uses NACCESS to compute which residues of both the receptor and the ligand are exposed at the surface of the proteins, and which are buried. This step is important to calculate the interaction surface between the receptor and the ligand. We use this information to evaluate the shape complementarity of the complex using discretized protein models. Subsequently, we compute the statistic potential: the propensity of different amino acids pairs in the interface to be in contact among resolved complexes. These propensities were computed with the Glaser method. Finally, we estimate the electrostatic energy and Lennard-Jones potential on the interaction site. Lennard-Jones potential takes into account Van der Waals forces.

To generate the machine learning algorithm, we used scikit-learn with a random forest model. We recovered the sampling data available in GitHub and data from another protein docking benchmark, making a total of four different natures of protein-protein interactions. Using the sampling script from team 6, we obtained 5936 complexes that were divided in two groups to be used as a training set for our machine learning algorithm. This allowed us to later score new complexes using a predicted TM score. Thanks to this scoring program, we were able to predict poses relatively close to the native one, on the 2PCC complex (fig2).

Our strategy is summarized in Figure 1.

Complementary strategy:

The complementary sampling strategy we chose is the one developed by team 6 because it performs a naïve sampling. It is the most appropriated for us because we already had implemented a shape complementarity method in our approach.

Added value / Opportunities of improvement:

The advantage provided by our scoring system is that we use a large amount of physico-chemical data to infer our score. This allows us to enrich the input information, increasing the robustness of our program. On the other hand, the use of a machine learning program to infer our final score increases the reliability of our results. In addition to all this, we have managed to minimize the running time of the program, which is why we consider it to be quite optimal.

Finally, we have several improvement proposals for our program. We would like to estimate more data at the beginning by training our machine learning on more complexes. A possibility could be to recover sets of known complexes whose sequences are similar to train our machine learning algorithm.

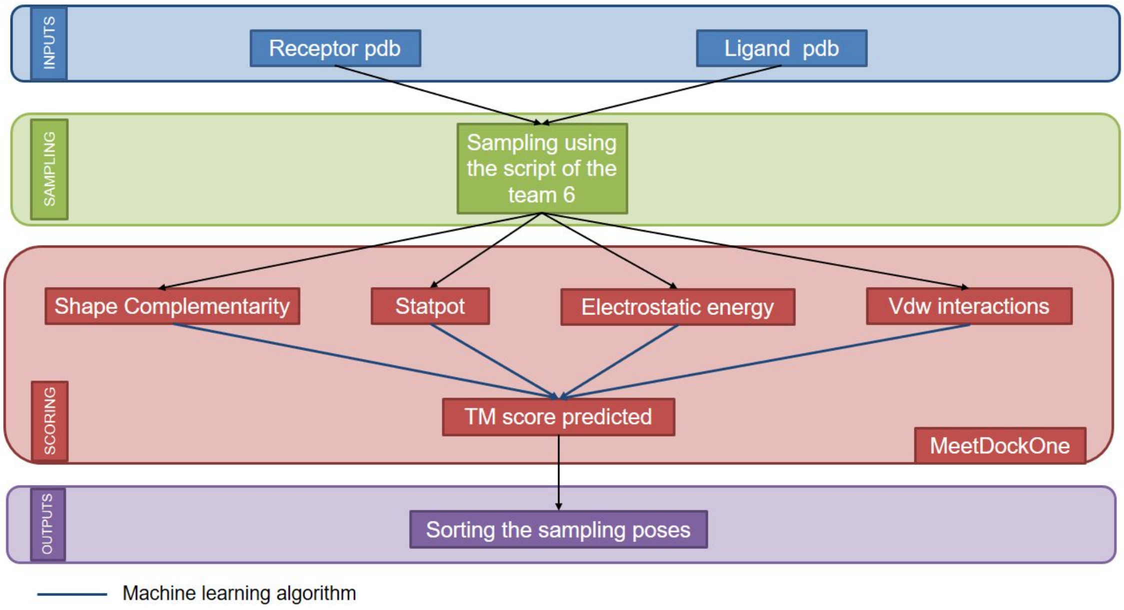

Figure 1: Pipeline of our whole docking program:

Abbreviations **Vdw:** Van der Walls, **Statpot:** Statistic potential, which studies the propensity of the amino acids in the interface to be in contact among resolved complexes.

Our scoring function MeetDockOne is summaries into the red rectangle.

Our output is double, a MeetDockOne.csv file which contains all the poses sorted regarding the Tm score predicted associated to a MeetDockOne.png which illustrated the scores distribution of all MeetDockOne functions and the predicted Tm score.

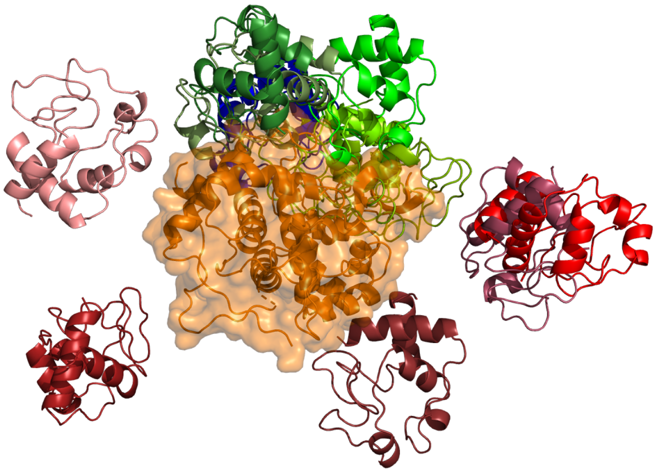

Figure 2: 2PCC native complex (receptor in orange, ligand in blue). Top 5 poses (shades of green), worst 5 poses (shades of red), according to MeetDockOne predicted Tm score