HADDOCK screening against the SARS-CoV-2 main protease (Mpro - 3CLpro)

P. I. Koukos, M. Réau, A. M. J. J. Bonvin*

July 2, 2020

Computational Structural Biology group, Bijvoet Centre for Biomolecular Research, Utrecht University, Netherlands

$*$: **a.m.j.j.bonvin@uu.nl**

Disclaimer

The compound rankings and binding poses presented in this work are the result of a single virtual assay and should not be considered meaningful until they have been experimentally validated.

Introduction

The novel coronavirus (SARS-CoV-2) that has emerged from Wuhan, China in December 2019 has spread to almost all countries in the world causing a dramatic number of deaths. The current absence of antiviral treatment against the SARS-CoV-2 urges the scientific community to accelerate the drug discovery research process.

One way to identify potential treatments and to be able to administer it swiftly is to focus on drug repurposing studies, i.e. to investigate the SARS-CoV-2 antiviral potential of drugs that have already been approved for human use.

Screening

Target

Proteins that are crucial for the survival and replication of the virus are the most attractive targets for such studies. Here we have focused on the SARS-CoV-2 main protease (3CLpro) that plays an essential role in the virus replication process by screening ~2000 approved drugs (and 6 experimental ones) against this particular protein.

For the docking we used PDB entry 6Y2F (Zhang et al. *Science*, 2020).

Dataset

Our dataset consists of the approved subset of the DRUGBANK database with additional filters for size and molecular weight, selecting only compounds whose weight is up to 750 g/mol and no less than 5 heavy atoms. In addition to these approved compounds we have added some experimental compounds which have been the focus of a plethora of scientific studies recently.

The list of compounds can be found in the `compounds.txt` file located in the same folder as this document. The compounds whose id starts with `DB` were extracted from the DRUGBANK database and the ones whose id starts with `CID` were downloaded from PubChem. These are all the active metabolites of the DRUGBANK compounds we could identify in public databases.

3D conformers were generated using OpenEye Omega (Hawkins *et al.* J. Chem. Inf. Model. *50* 572-584 (2010)).

Protocol

The rationale behind HADDOCK is to make use of experimental information to guide the docking. Herein, we took advantage of the large amount of high quality *holo* structures of the SARS-CoV-2 3CLpro and related proteins ($> 90\%$ identity) published in the Protein Data Bank. Among those crystallographic data, 66 molecules are non-covalent and covalent active-site fragments from the large XChem crystallographic fragment screen against 3CLpro performed by Diamond. In total, we collected 92 molecules targeting the 3CLpro(-related) binding site.

We identified one crystallographic template to be used for the docking of every target compound in the virtual library. For the selection, we calculated the Tanimoto coefficient computed over the Maximum Common Substructure (MCS) between target and template compounds and selected the template with the highest Tanimoto coefficient. After superimposing all templates we transformed the heavy atoms of their compounds into shape beads and defined restraints between the beads and heavy atoms of the target compound to guide it in the binding pocket.

The compounds were scored using the HADDOCK scoring function and the scores are reported in in arbitrary units:

```
HADDOCKscore =  1.0 Evdw + 0.1 Eelec + 1.0 Edesol
```

where:

- *Evdw* is the van der Waals intermolecular energy
- *Eelec* is the electrostatic intermolecular energy
- *Edesol* is an empirical desolvation energy term.

The resulting models were structured using an RMSD cutoff of 1.5Å and a minimum cluster membership of 4. In the cases were no clusters could be formed with these criteria, the cluster membership requirement was lowered to 2 and if clusters still couldn't be formed the models were ranked individually. For all other cases clusters were ranked using the HADDOCK score of the top compound of every cluster. The score of the top model of the top cluster is the one that is presented in the final table.

Results

File `summary.txt` contains a table which summarises the main findings of this virtual screening assay. Every compound is associated with its name and ATC codes, one of four categories depending on its activity and a classification depending on whether it is approved or not. The 4 categories are 'Protease Inhibitors', 'Antivirals', 'Antiinfectives' and 'Other'. The first is made up compounds that are known to inhibit proteases, the second antiviral medications, the third general antiinfectives and the last everything else. The 'NA' group corresponds to compounds that have no specific associations - these tend to be things like dietary supplements, aminoacid residues, etc… The `summary.txt` file is sorted by target id with the DRUGBANK targets appearing first, followed by the PubChem ones. The file `summary-haddock_score-sorted.txt` ranks the targets by HADDOCK score (see above).

Interactive versions of this table along with visualisation of the overall results can be found on our website. The tables and graphs for the Mpro screening collate the results of two assays: The one presented in this document and another based on pharmacophore similarity.

The top poses of the top 20 compounds, according to the joint ranking presented on the bonvinlab.org website, are being studied with Molecular Dynamics (MD) simulations as well.

In addition to this summary the full results for every target are made available in the `results` folder which can be found in the directory as this document. It contains one folder per entry of the dataset (excluding the ones for which pharmacophore features could not be calculated and were, therefore, not included), using the compound id as the folder name. In those folders one can find PDB anywhere from 20 to 400 PDB files, depending on the number of conformers that were generated and some accompanying files. These are:

```
* clusters.stat
* clusters.stat_best1
* file.list
* file.list_clustX
* file.list_clustX_best1
```

where X stands for the cluster number. Clusters are numbered according to their size, ie cluster 1 is going to be the most sizeable followed by clusters 2, 3 and so on. The file `clusters.stat` offers statistics on all clusters using various metrics which are listed in the header of the file. The file `clusters.stat_best1` reports the same value but calculated only over the top model of every cluster instead of averaged over the entire cluster. The `file.list_clustX` and `file.list_clustX_best1` files report the scores of every model and the top model of every cluster, respectively. `file.list` lists the scores for all models without taking clustering into consideration.

Acknowledgements

Dr. Ed Moret from the Department of Pharmaceutical Sciences at Utrecht University has provided valuable expertise on the compounds and their relevance.