Molecular docking by MolAICal

Qifeng Bai
Email: molaical@yeah.net
Homepage: https://molaical.github.io or https://molaical.gitlab.io
School of Basic Medical Sciences
Lanzhou University
Lanzhou, Gansu 730000, P. R. China

1. Introduction

SARS-CoV-2 caused the rapid spread of coronavirus disease 2019 (COVID-19) throughout the world. In this tutorial, the SARS-CoV-2 main protease (Mpro) which plays an important role in the replication of coronavirus is selected as the example target. The crystal structures of SARS-CoV-2 Mpro have been reported (PDB ID: 6LU7, 6Y2F, etc) [1, 2]. In this tutorial, the molecular docking between protein and ligand is introduced based on MolAICal (https://molaical.github.io). Autodock Vina has the Pearson and Spearman correlation coefficients (rp/rs) are 0.5259 and 0.5421 based on the experimental binding affinity of the 3130 complexes if the ligand with lowest RMSD has the lowest binding affinity value among the 20 docked ligands. For MolAICal, rp/rs are 0.5335 and 0.5489 at the same assay conditions of Autodock Vina. It indicates that MolAICal has better 'docking' and 'ranking' power than Autodock Vina.

2. Materials

2.1. Software requirement

1) MolAICal: https://molaical.github.io

2) UCSF Chimera: https://www.cgl.ucsf.edu/chimera

2.2. Example files

1) All the necessary tutorial files are downloaded from:
https://github.com/MolAICal/tutorials/tree/master/0000-docking

3. Procedure

3.1. Dealing with receptor and ligand

1. Open the file "6y2f.pdb" of SARS-CoV-2 main protease in complex with ligand (PDB ID: 6Y2F): File→Open (see Figure 1).

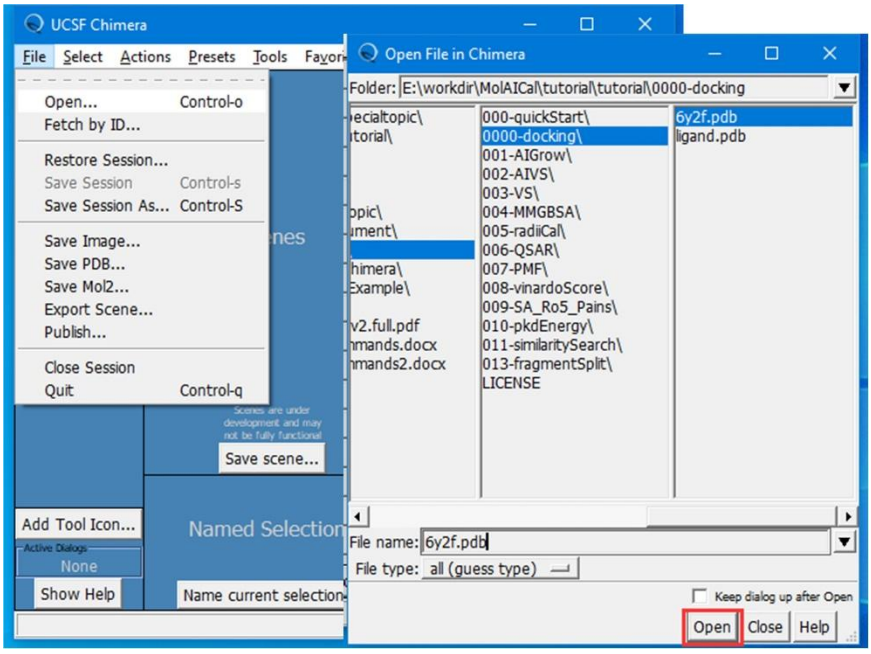

Figure 1

2. Prepare the Mpro receptor and save Mpro receptor named “protein.pdb”. The detail procedure is shown in the Figure 2.

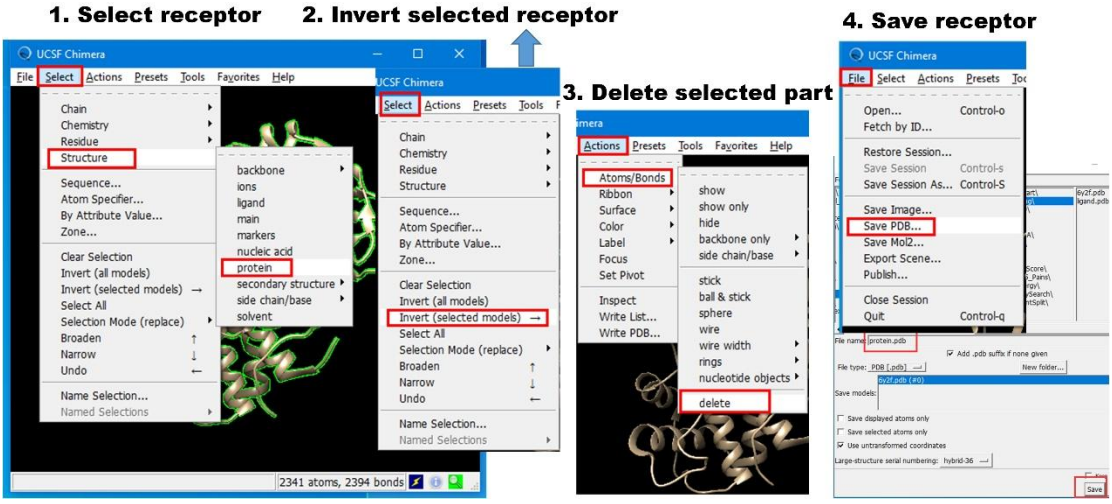

Figure 2

3. The ligand has been deleted with the above steps, users can choose to close UCSF Chimera and re-open file “6y2f.pdb” as above steps (**Or** click “File→Close Session” based on the above UCSF Chimera state to close the current session and reload file “6y2f.pdb”). Now, prepare the ligand and save the ligand named “ligand.pdb”.

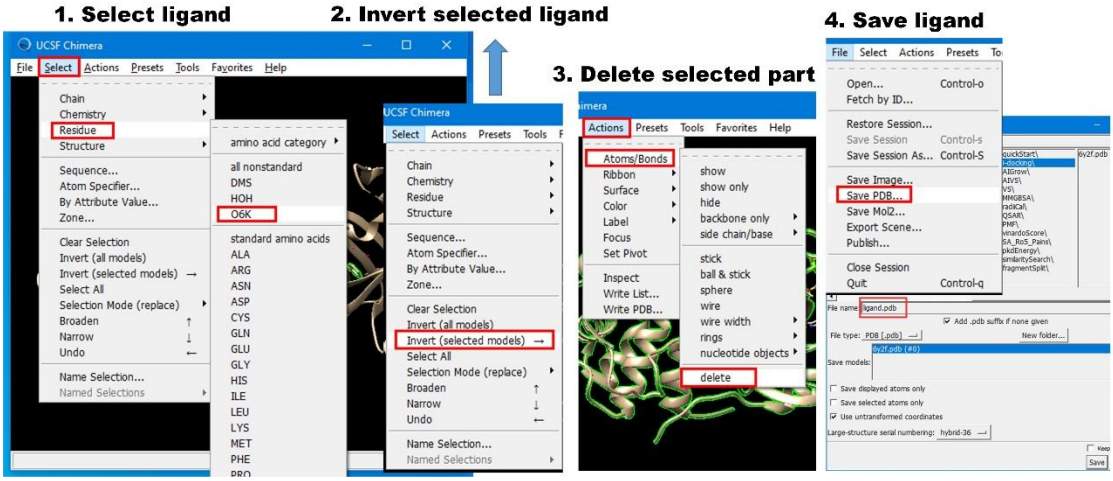

Figure 3

3.2. Convert receptor and ligand into PQBQT format

1. Get PDBQT format of receptor by using below command:

#> MolAICal-xxx\molaical.exe -dock receptor -i protein.pdb

Note: MolAICal-xxx is your downloaded version of MolAICal.

It will generate the file named “protein.pdbqt” which has the same prefix name of “protein.pdb”.

2. Get PDBQT format of ligand by using below command:
#> MolAICal-xxx\molaical.exe -dock ligand -i ligand.pdb

It will generate the file named “ligand.pdbqt” which has the same prefix name of “ligand.pdb”.

Notice: Ligand should contain full structure. If ligand is **lack of hydrogen**, MolAICal may not generate ligand in PDBQT format. Users should firstly employ UCSF Chimera to add hydrogen on the ligand which is lack of hydrogen, or use MolAICal to convert ligand in PDBQT format by the following command (users should replace 1.mol2 into their own filename):
#> molaical.exe -tool format -i E:/1.mol2 -o E:/1.pdbqt

3.3. Obtaining the center and length of docking box

1. Open protein.pdb and ligand.pdb in order. And then open “Command Line” of Chimera: Favorites→Command Line (see Figure 4).

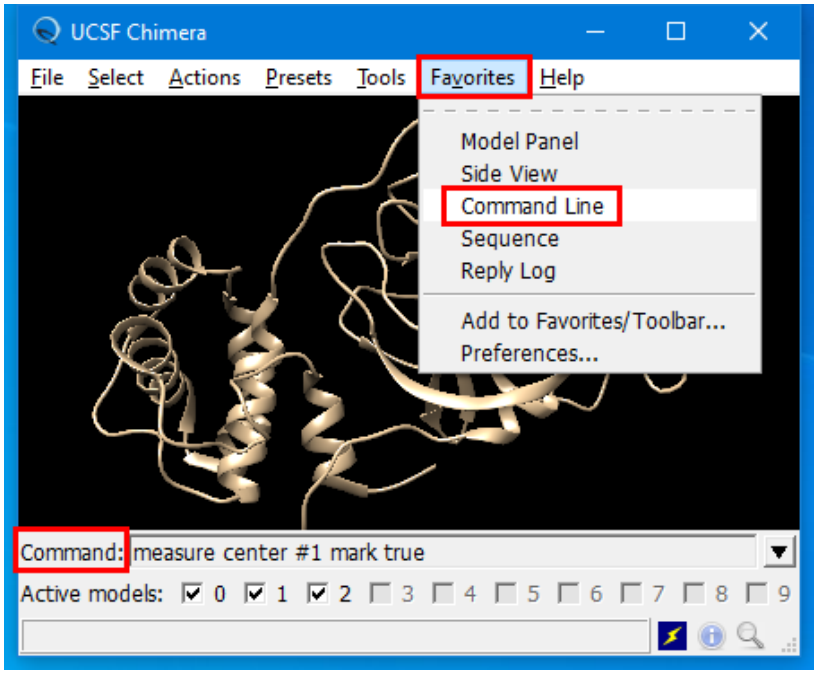

Figure 4

2. Make sure the open sequence of ligand. If protein is open firstly, it will correspond to “Active models 0”. Second open corresponds to “Active models 1”, and so on (see Figure 5). Here, ligand is secondly open (“Active models 1”). Put the below command in to command line (see Figure 5):
define centroid mass false #1
And click **“Enter” key. And then, click following the sequences in Figure 5.** It will show geometric center coordinates (x, y, z) of ligand is (10.879, -0.251, 20.754).

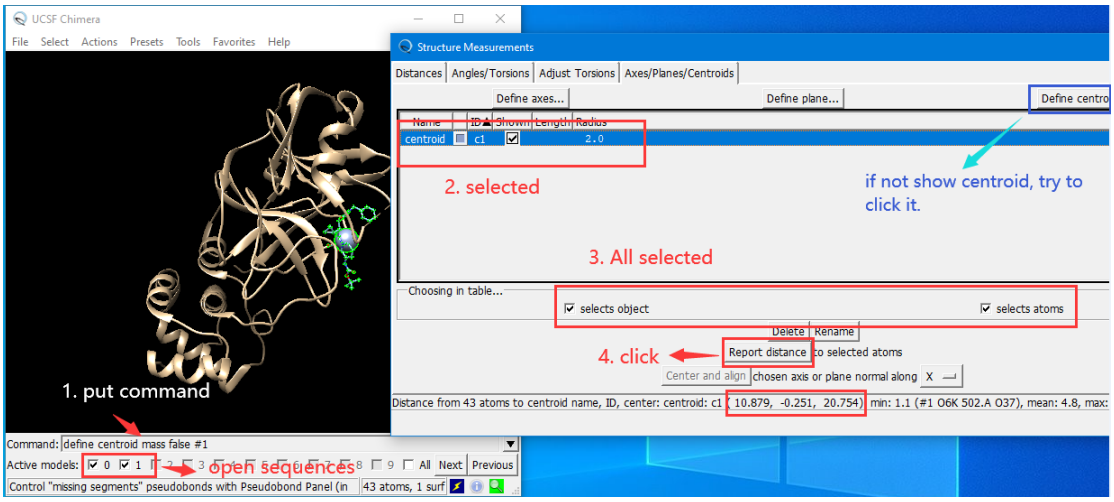

Figure 5

3. Determine the box size and center by UCSF chimera.

Open box tool: Tools→Surface/Binding Analysis→Autodock Vina

Check box: Select the right receptor (here, it is named "protein.pdb") and ligand (here, it is named "ligand.pdb") (see Figure 6). Put the above center coordinates "10.879, -0.251, 20.754" into center box (see Figure 6), the size can be tried by users until the box has a suitable size.

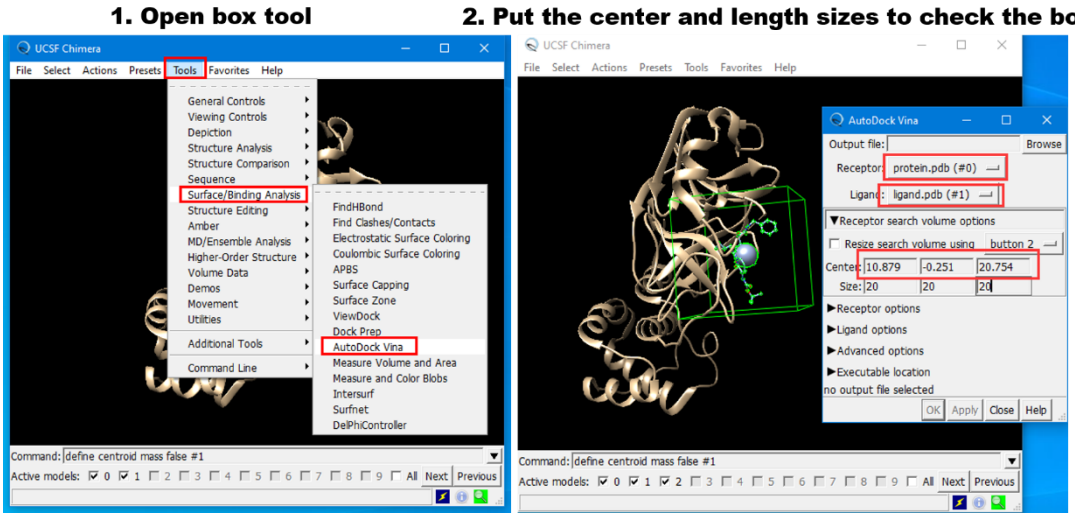

Figure 6

Note: the users can tick "Resize search volume using button 1, 2 or 3". Button 1, 2 or 3 represents left, middle or right click of mouse. If the users select this function, they adjust the box size via mouse. If you are interested in it, you can try this function.

4. Assuming the configure file is named "conf.txt", the final configure file can be written as follows:

```
out = all.pdbqt
cpu = 4
receptor = protein.pdbqt
center_x = 10.879
center_y = -0.251
```

center_z = 20.754
size_x = 20
size_y = 20
size_z = 20
num_modes = 3

Where “Out” is the output file name, “cpu” is the number of the used CPU cores, “receptor” represents the receptor name, and “num_modes” is the number of generated docking conformations. If “num_modes” is 3, it will generate 3 docking structures of the ligand.

3.4. Molecular docking by MolAICal

MolAICalD has now been fully integrated into MolAICal, allowing users to perform molecular docking directly through “**molaical.exe**” without the need to explicitly use “molaicald” binary executable for molecular docking between receptor and ligand. The command is below:
#>MolAICal-xxx/molaical.exe -dock sd -r protein.pdb -i --config conf.txt --ligand ligand.pdbqt

Note: “MolAICal-xxx” represents the actual installation directory of MolAICal. The parameters "-dock" and "-r" are built-in parameters of MolAICal, while the parameter following "-i" belongs to the molecular docking module MolAICalD. **Mandatory requirement: "-dock" and "-r" must be placed before "-i".**

It will automatically **generate molecular files prefixed with numbers, including '.mol2', '.pdbqt', and '_min.mol2' formats**. Specifically:

- ✧ **Number-prefixed '.pdbqt' files** represent individual conformation molecules in PDBQT format after splitting molecular docking results. For example, the file name is “**1.pdbqt**”.
- ✧ **Number-prefixed '.mol2' files** represent individual conformation molecules converted to MOL2 format after splitting molecular docking results. For example, the file name is “**1.mol2**”.
- ✧ **Number-prefixed '_min.mol2' files** represent the energy-minimized molecules based on the docking results. For example, the file name is “**1_min.mol2**”.

Energy minimization helps resolve excessive ligand distortion issues and optimizes post-docking molecules. Users can select appropriate molecules based on research requirements: for example, choosing either the docked molecules alone or the energy-minimized molecules for further analysis.

Options: If the user **does not wish to perform post-docking processing (such as splitting results or energy minimization),** the '-r' parameter should be omitted. In this case, the program will only execute molecular docking without any subsequent operations. The specific command is as follows:
#>MolAICal-xxx/molaical.exe -dock sd -i --config conf.txt --ligand ligand.pdbqt

Sometimes, **the ligand, such as this ligand in this tutorial,** has many rotatable bonds; under these circumstances, the molecular docking should be run many times to compare with the original crystal ligand. And then, users can screen the suitable random seed to repeat the good results, and use this random seed for further molecular docking and virtual screening (see Figure 7).

For example, I screen random seed 555767984 for better docking results (see Figure 7). Users can repeat the better results like me by using the random seed of 555767984. Please input the command below:

#>MolAICal-xxx/molaical.exe -dock sd -r protein.pdb -i --config conf.txt --ligand ligand.pdbqt --seed 555767984

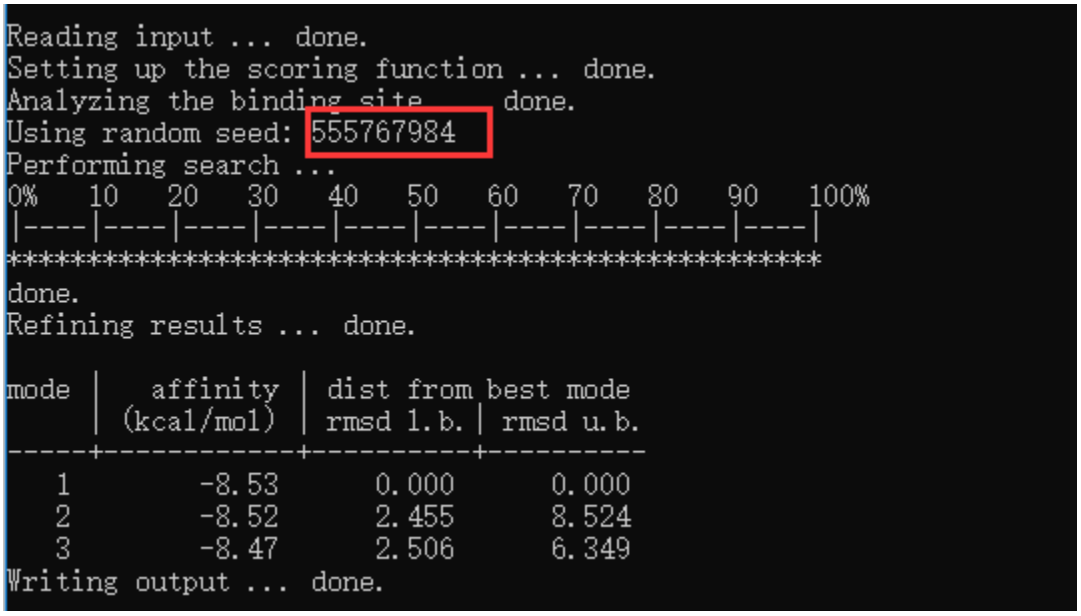

Figure 7

The single molecule is named 1.pdbqt (or **1.mol2**), 2.pdbqt (or **2.mol2**), 3.pdbqt (or **3.mol2**), or **all.pdbqt**, *etc*. "1.pdbqt (or **1.mol2**)" contains the docking conformation with the best binding affinity and the like.

Users can check the 1.pdbqt, 2.pdbqt, or 3.pdbqt by **Pymol** (https://github.com/cgohlke/pymol-open-source-wheels) or **ChimeraX** (https://www.cgl.ucsf.edu/chimerax) directly.

Here, UCSF Chimera is used to check results in mol2 format. Now, open UCSF Chimera and load "**protein.pdb, 1.mol2, 2.mol2, and 3.mol2**", *etc*. Users can choose to show or hide molecules when all molecules are loaded via Favorites→Model Panel (see Figure 8).

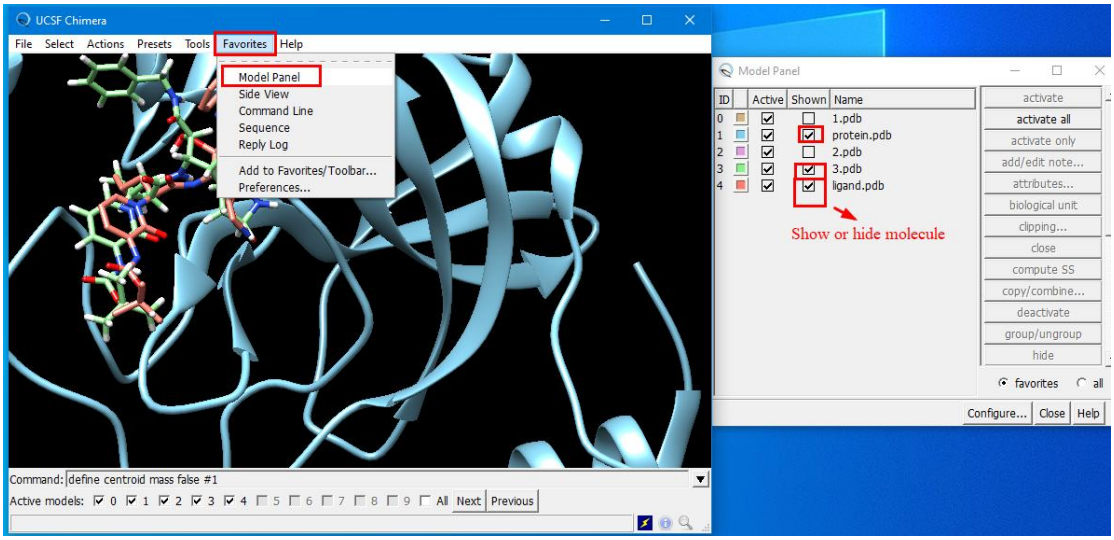

Figure 8

Note: Users can choose to add hydrogen to "protein.pdb" if they want to analyze the interaction between the protein and the docked ligands.

The results show "1.mol2" has some part overlap with the original crystal ligand, while "3.mol2" has a similar pose to the original ligand. I had performed molecular dynamics (MD) simulations on this system (**see: MM/GBSA tutorial in https://molaical.github.io/tutorial.html**). The MM/GBSA results based on MD simulations show Andricioaei entropy of original crystal ligand N3 is -84.70646297459386 (kcal/mol). It indicates the crystal ligand N3 is not stable in the pocket of SARS-CoV-2 main protease.

3.5. Analysis: showing the hydrogen bond interaction between protein and ligand.

This part is optional. If users want to analyze the interaction, such as the hydrogen bond between receptor and ligand, they can refer to this part in this tutorial.

1) I assume hydrogen atoms have been added to "protein.pdb" and "1.pdb". Open "protein.pdb" and "1.pdb" in the same window of UCSF Chimera. And then, open "Tools→Surface/Binding Analysis→FindHBond" (see Figure 9). Following the steps in Figure 9, it may show hydrogen bonding interaction between partial residues of the receptor and ligand. (**Notice:** it shows many functions in Tools→Surface/Binding Analysis, users can choose any function for their special analysis purpose. Here, I only show the hydrogen bonding interaction analysis between receptor and ligand.)

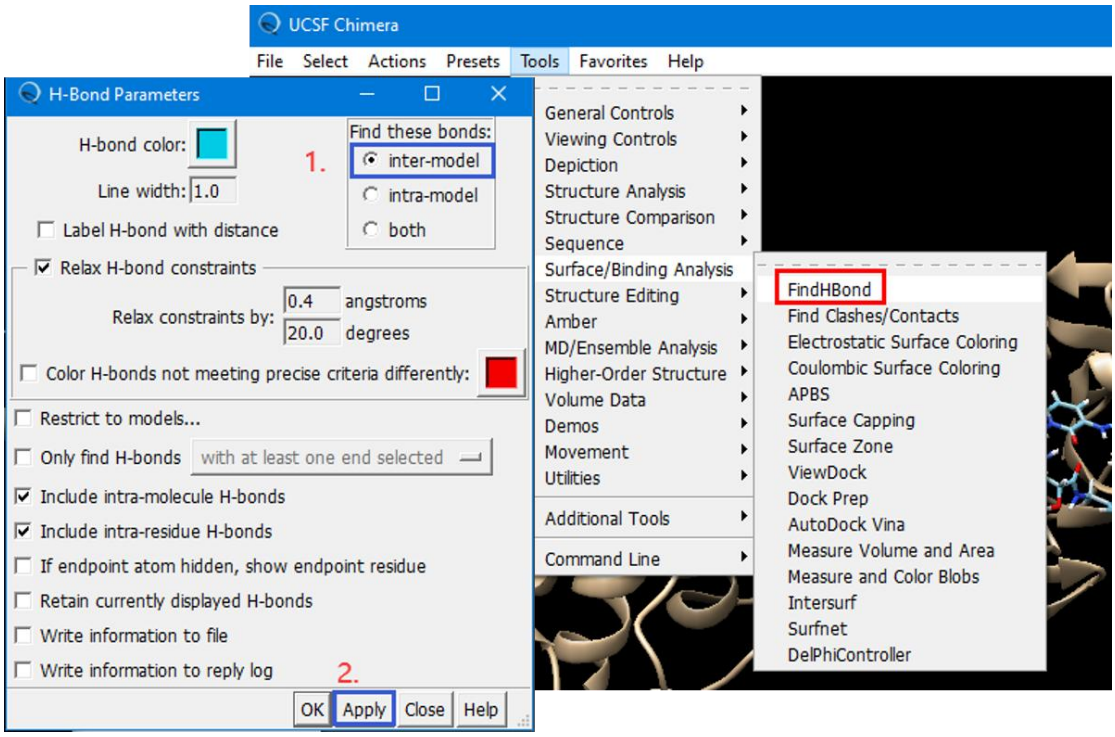

Figure 9

2) To show all residues of the receptor that interact with the ligand, it should write information to a file. Please follow the steps in Figure 10. It will generate an information file named “hbond.info” which contains hydrogen bonding information as follows:

#0 SER 144.A OG #1 O6K 502.A O48 no hydrogen 2.817 N/A
#0 CYS 145.A SG #1 O6K 502.A O48 no hydrogen 3.160 N/A
#0 GLU 166.A N #1 O6K 502.A O37 no hydrogen 2.933 N/A

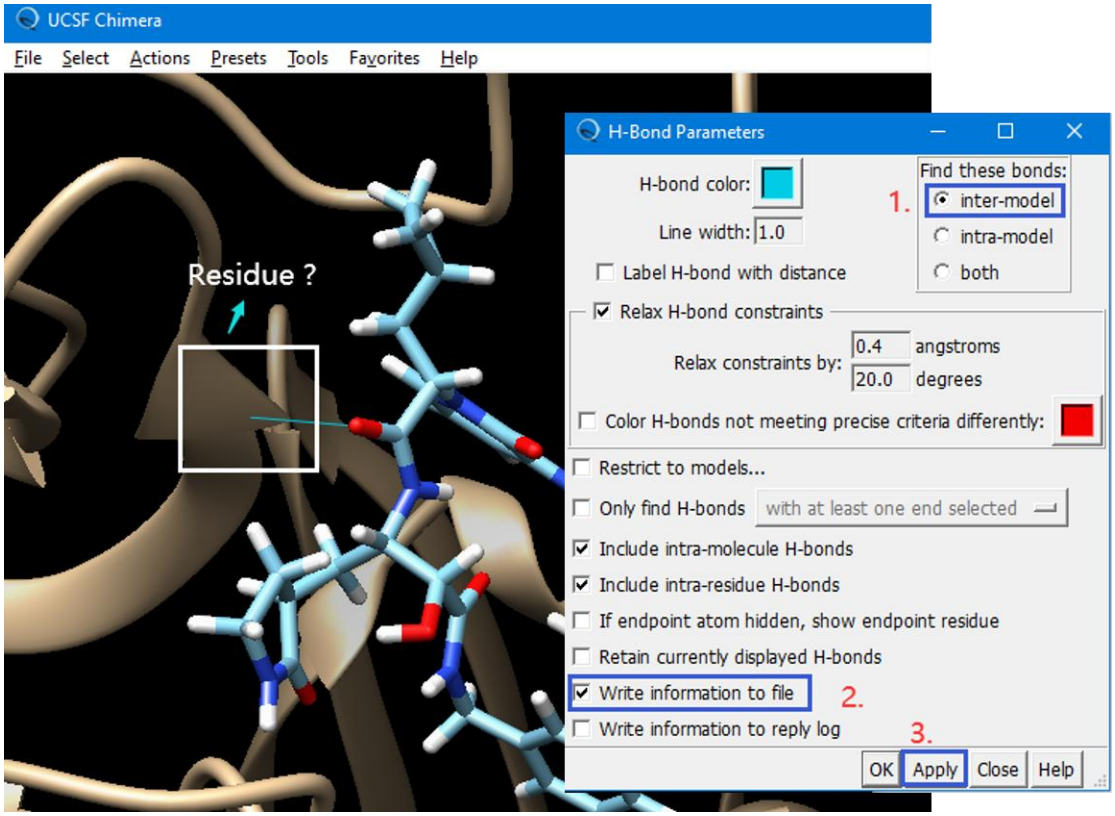

Figure 10

3) Select the residues (Here, they are SER144, CYS145, and GLU166) in "hbond.info", and show hydrogen bonding as follows in Figure 11:

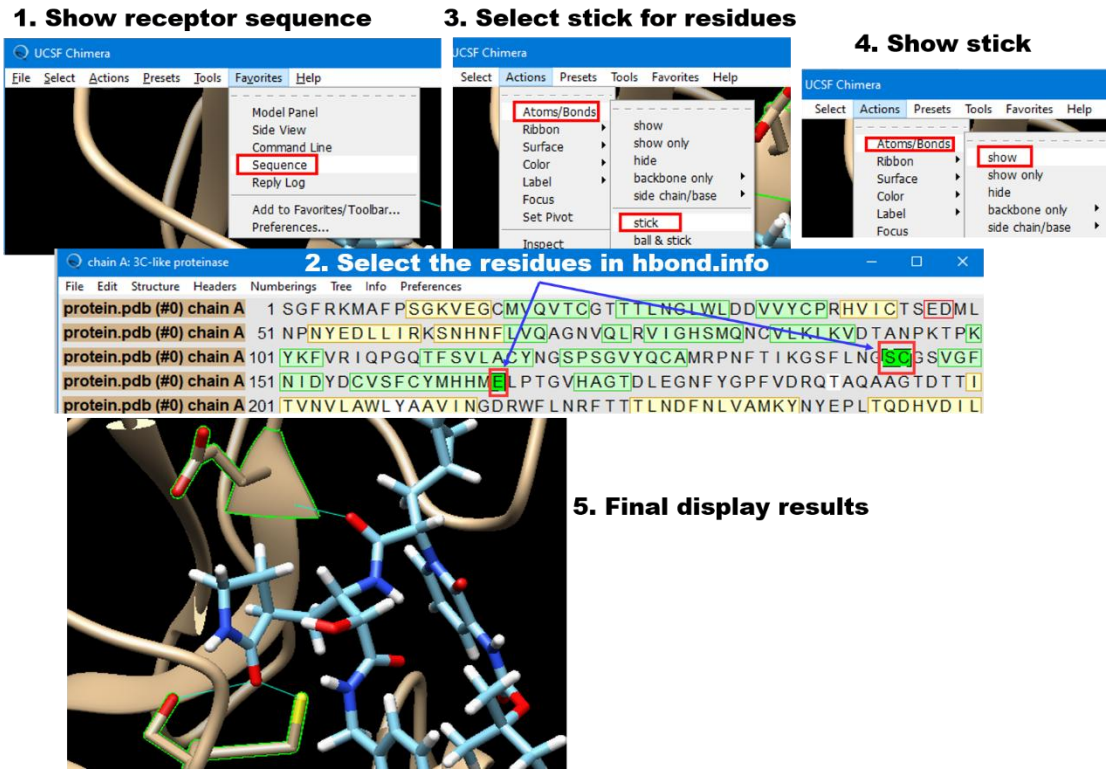

Figure 11

Tricks

If users want to obtain the publication style for displaying beautiful results, users can follow the steps in Figure 12. Finally, the display result is saved to an image named “displayResult.png”. Please check it. Of course, users can choose their desired style for the final display results.

Notice: It should select publication mode before hydrogen bonding analysis. If not, the stick style of residues will be missed when switching publication mode.

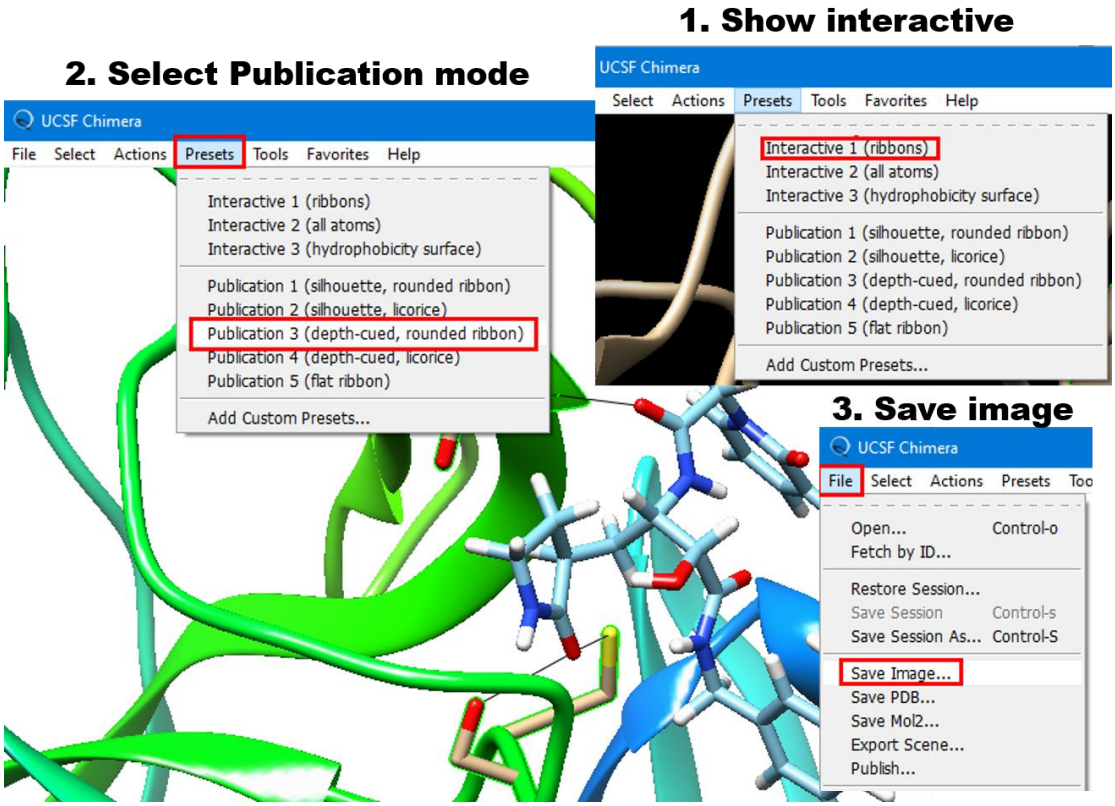

Figure 12

Appendix

If the user **does not perform post-docking processing (such as splitting results or energy minimization)** without setting the '-r' parameter as follows:

#>MolAICal-xxx/molaical.exe -dock sd -i --config conf.txt --ligand ligand.pdbqt

And then the **post-docking processing (such as splitting results or energy minimization) is required.** The commands below provide a good way to achieve this goal:

1. Splitting results into single molecule

#> MolAICal-xxx\molaical.exe -tool pdbqt -i all.pdbqt -o ./

The single molecule is named 1.pdbqt, 2.pdbqt, 3.pdbqt, or **all.pdbqt**, *etc*. "1.pdbqt" contains the docking conformation with the best binding affinity and the like.
Users can check the 1.pdbqt, 2.pdbqt, or 3.pdbqt by **Pymol** (https://github.com/cgohlke/pymol-open-source-wheels) or **ChimeraX** (https://www.cgl.ucsf.edu/chimerax) directly. Here, UCSF Chimera is used to check results.

2. For UCSF Chimera to load molecules correctly, the "pdbqt" format must first be converted to "pdb" or "mol2" format using MolAICal through one of the following two options:

Option 1:

1) To make sure the same way for adding hydrogen, first delete all hydrogen, and then add hydrogen

#> MolAICal-xxx/molaical.exe -tool format -i 1.pdbqt -o 1.mol2 -c delh
#> MolAICal-xxx/molaical.exe -tool format -i 1.mol2 -o 1.mol2 -c addh

Option 2:

1) Adding hydrogen (option)

#> MolAICal-xxx\molaical.exe -dock addh -i 1.pdbqt

2) Changing "pdbqt" to "pdb" format

#> MolAICal-xxx\molaical.exe -dock pdbqt2pdb -i 1.pdbqt

Users can use the same way for 2.pdbqt and 3.pdbqt in this tutorial. Now, it can open UCSF Chimera and load the "protein.pdb, 1.mol2 (or 1.pdb), 2.mol2 (or 2.pdb), and 3.mol2 (or 3.pdb)" for further check.

3. Perform energy minimization of the docked ligands, run the minimization command below:

#> MolAICal-xxx\molaical.exe -min2 -m 1 -i1 protein.pdb -i2 1.mol2 -o2 1_min.mol2 -f1 protein.pdb -f2 0

Note: the '-f2' is set to 0, which means no fixed atom in the second ligand named "1.mol2". For more detailed references on minimization, please check the MolAICal manual.

When users select Option 2 (PDB files only), the generated PDB files will contain formatting issues with hydrogen atoms. Consequently, these **hydrogen atoms** must first be **removed** and then properly **re-added** to the PDB files (e.g., 1.pdb) by Pymol or UCSF Chimera. Once the hydrogen atoms of the ligand are handled, the following command can be used for energy minimization based PDB-format file:

#> MolAICal-xxx\molaical.exe -min2 -m 1 -i1 protein.pdb -i2 1.pdb -o2 1_min.mol2 -f1 protein.pdb -f2 0 -c2

Note: PDB-format molecular files often lack covalent bond connectivity information. During energy minimization, this may lead to incorrect bond assignments. Therefore, it is necessary to convert PDB files into Mol2-format files, which include proper covalent bond connectivity data. The '-c2' means 1.pdb is first converted into a mol2-format file, then it is performed for energy minimization. For more detailed references on minimization, please check the MolAICal manual.

References

1. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. bioRxiv. 2020.
2. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. Science. 2020. doi: 10.1126/science.abb3405. PubMed PMID: 32198291.