Protein-Protein Docking Tutorial

Computational protein-protein docking is the process of predicting the structure of a protein complex from the structures of each independent component. These techniques can generally be divided into global docking, where the binding interfaces are completely unknown, or local docking, where a rough idea of where the two systems come into contact are known.

This tutorial presents a cross-docking benchmark experiment. The antibody CR6261 binds to multiple sub-types of influenza antigen hemagglutinin (HA). It has been crystallized with both H1 and H5 HA sub-types. In this tutorial the antibody from one crystal structure (3GBN) will be docked to the antigen from the other crystal structure (3GBM).

Initial set-up

Hemagglutinin is normally a trimer, but for this tutorial we'll be docking it as a monomer. (The antibodies we're working with bind only to a single monomer.)

1. Create a new working directory, and download the 3GBN and 3GBM structures as PDB format.
    1. Go to https://rcsb.org and type '3gbn' in the search bar.
    2. Click on 'Download Files' on the right side of the page, then 'PDB Format'.
    3. Save the PDB file in your working directory as `3gbn.pdb`.
       (Note, the file may automatically download to your "Downloads" folder. If so, move the file into your working directory.)
    4. Repeat for 3gbm
2. Open both structures in ChimeraX, and examine the chains for each structures. For structure 3GBM we have two complexes in the same crystal unit: ABHL and CDIM. The chains AB/CD represents the head domain and the stalk domain of the hemagglutinin protein, and HL/IM the CR6261 heavy and light chains. For structure 3GBN we have the single complex ABHL. In the next step we will extract the chain of interest for this cross-docking experiment: chains AB for 3GBM and chains HL for 3GBN.

Structure pre-processing

We need to extract only the chains of interest from each structure. With both structure open in ChimeraX:

1. From 3GBM we want chains A & B, but only the protein residues
    1. Select the chains: `select #1/A,B & protein` (assumes 3GBM is model #1)
    2. Save the chains as 3gbm_AB.pdb
       (File->Save->File of Type->PDB->Save selected atoms only)
2. From 3GBN we want to save chains H & L. But when you examine the structure, notice how the H&L chains contains incomplete portions of the full-length antibody. For the purposes of docking, we only need the Fv region (the well-formed domains).
    1. Visualize the C-terminal bit of chain H to remove: `color #2/H:112-188 magenta` (assumes 3GBN is model #2)
    2. Remove the same selection: `delete #2/H:112-188`
    3. Visualize the C-terminal bit of chain L to remove: `color #2/L:106-195 magenta`
    4. Remove the same selection: `delete #2/L:106-195`
    5. Select the chains: `select #2/H,L`
    6. Save the file as 3gbn_HL.pdb
       (File->Save->File of Type->PDB->Save selected atoms only)

Reopen the 3gbm_AB.pdb and 3gbn_HL.pdb structures in ChimeraX, making sure the structures are what you expect them to be.

Refine the starting structures

Structure refinement is essential for any post-processing step in Rosetta to remove clashes that might be present in the crystal structure. Refinement can be performed at multiple levels: repack (side-chains only), minimization (side-chains and backbone) or relax, which involve multiple cycles of repack and minimization.

We will use the ROSIE relax application to relax the structures into the Rosetta energy function. To keep the structure backbones close to what was crystallized, we'll be adding coordinates tethers (restraints/constraints) to the backbone atoms.

1. Go to https://rosie.rosettacommons.org and click the Relax application.
2. Log in with your Github account, if prompted.
3. Upload the 3gbn_HL.pdb structure on the submission page
4. Check the box labeled "tether backbone coordinates of the pdbs being relaxed to the coordinates in the xtal native". Don't tether sidechain atoms
5. Give the job a descriptive name (e.g. "Relax 3gbn_HL.pdb") and press "upload and queue job"
6. Theoretically, you should repeat the submission for the 3gbm_AB.pdb structure, but as it's larger, it takes much longer to relax. As such, we've provided a pre-relaxed version of the structure for you to use.

The relax job will likely take some time (5-10 minutes) to make it to the front of the queue and run. You can keep the post-submission browser tab open to monitor the status, or go to https://r2.graylab.jhu.edu/queue to check the state of your jobs.

- Once the job completes, go to the results page and click the download button next to "output/relaxed.pdb" to download the structure. Save the file as 3gbn_HL__relaxed.pdb (or 3gbm_AB__relaxed.pdb) in your working directory.

It can also be useful to pre-generate backbone conformational diversity prior to docking particularly when the partners are crystallized separately. In this case, backbone tethers aren't used, and multiple relaxed outputs with diverse backbone structures are docked with each other in separate docking runs. However, backbone conformational diversity will not be explored in this tutorial due to time limitations.

Protein-protein docking setup.

For Rosetta protein-protein docking, both partners need to be in the same input PDB. It's also ideal if the structures are roughly in their docked conformation (local docking). We can use this information to improve the efficiency of the docking process and the quality of the final model.

For our docking protocol, we already know where the structures should roughly be docked, as we have a co-crystal structure of the partners.

1. Open 3gbn.pdb as well as the 3gbm_AB__relaxed.pdb and 3gbn_HL__relaxed.pdb structures in ChimeraX.

   The following assumes 3gbn.pdb is model #1; 3gbm_AB__relaxed.pdb is model #2 and 3gbn_HL__relaxed.pdb is model #3

2. Align the H & L chains of 3gbn_HL__relaxed.pdb to the H & L chains of 3gbn.pdb (via CA atoms)

   ```
   align #3/H,L@CA & protein toAtoms #1/H,L@CA & protein matchNumbering true
   ```

3. Align the A & B chains of 3gbm_AB__relaxed.pdb to the A & B chains of 3gbn.pdb

   ```
   align #2/A,B@CA & protein toAtoms #1/A,B@CA & protein matchNumbering true
   ```

4. Save the re-aligned chains to a file `combined.pdb`

   File->Save->File of Type->PDB->(Make sure only 3gbm_AB__relaxed.pdb and 3gbn_HL__relaxed.pdb are selected in the box – use ctrl-click or shift-click to select both.)

5. Close/hide all open structures and open the combined.pdb file – check to make sure it's what you expect it to be, has the all the chains in the expected relative orientation, and does not contain any extra residues from 3gbn.pdb.

Docking

We'll be using the ROSIE Docking application. This application is built for general protein-protein docking, but can be used for protein-antibody docking.

1. Go to https://rosie.rosettacommons.org and click the Docking application.
2. Click the "[documentation]" link to read the Docking documentation.
3. Upload the combined.pdb file to the "Combined PDB file" entry.
4. Uncheck "use ensemble docking"
5. Leave the reference model box blank
6. Set the docking partners field to "AB_HL" (dock chains H&L as a unit to chains A&B)
7. Enter a descriptive job name and click the "Upload and queue job" button.

Protein-protein docking can take some time to run. You can work ahead with the post-analysis with the example structure provided.

Analysis of docking results.

Interpreting ROSIE output

The ROSIE output page presents the 10 top scoring structures by particular score metrics. The default is to rank the structures by the interface binding score (I_sc). This is Rosetta's best guess as to the binding energy (in arbitrary units), with more negative being better. This can be changed to other metrics, but there generally isn't a reason to do so.

To help interpret the docking results, it also presents a score-versus rmsd plot (where by default the rmsd is plotted to the input structure). This "funnel plot" should ideally show the low-energy structures as all being low rmsd. Due to the Rosetta energy function being rough, it's not a problem if a low rmsd structure is a higher energy. So as you go from high energies to lower energies, the plot for a successful run should make a "funnel" being broad at the top and narrow at the bottom, pointing toward 0 Ang rmsd. A successful funnel indicates that the Rosetta sampling method was able to successfully pick out a preferred (hopefully native) structure. While the stochastic nature of the Rosetta sampling algorithm means that not all output structures find the bottom of the docking well, having a number of structures which can pick it out consistently lends support to the idea that the low energy structures are native-like.

The calculated rmsd depends on the reference structure used. If your input structure isn't representative of the native structure, then the tip of the funnel may point to a higher rmsd value. Note, though, that rmsd space is not even. There's "more room" at higher rmsds, so while two structures at 1 Ang rmsd are likely close to each other, two structures both at 10 Ang may not be similar to each other. So to be safe, if the funnel is pointing to a high rmsd structure, you often want to replot with rmsds calculated with the low energy structure as a reference. (However, if all the low energy structures are similar to each other when examined in ChimeraX or the like, then it's likely the docking run is successful

Examining residue energy contributions.

ROSIE can provide energy information for protein structures (including protein-protein complexes), allowing a more detailed residue-by-residue examination of the interface.

1. Go to https://rosie.rosettacommons.org and click the "Energy Breakdown" application.
2. Upload the docked structure (if you're working ahead, use the files/results/r_0715.pdb structure provided.)
3. The default values of "ignore waters", "ref2015" and not parameter files should be used.
4. Enter a descriptive job name and click the "Upload and queue job" button.

When the job finishes, the main outputs will be `per_residue_energies.tsv` and `residue_energy_breakdown.tsv` scorefiles in the "outputs" section. Download these files.

The files are in a tabular format, and list calculated Rosetta Energy values for each residue (`per_residue_energies.tsv`) or for residue pairs (`residue_energy_breakdown.tsv`). These files can be opened with a spreadsheet program (e.g. Excel, LibreOffice Calc, Google Sheets). When importing, select the options to use spaces as delimiters and to merge adjacent delimiters.

The files list all energies, including residue-internal energies and intra-chain energies. For docking purposes, we're mainly interested in the inter-chain residue energies.i

1. Open `residue_energy_breakdown.tsv` with your spreadsheet program of choice.
2. Residue pairs should be listed with the partner which comes earlier in the PDB coming first. We can use that to select only those lines which are between the two chains.
    1. Highlight and delete all the "onebody" terms at the top of the sheet.
    2. Scroll down to where the `pdbid1` listing switches from one binding partner to the other (e.g. A/B to H/L, or vice versa).
    3. Highlight all the lines after the switchover point and delete them.
    4. Sort by the `resid2` column.
    5. Scroll down to where the `pdbid2` listing switches from one binding partner to the other (e.g. A/B to H/L, or vice versa).
    6. Highlight all the lines *before* the switchover point (except the column label line) and delete them.
    7. If successful, all the remaining lines should just be between residues in one partner and the other.
3. Sort the remaining lines (ascending) by the `total` column. More negative scores are better.

Residue-residue pairs with a highly negative score are contributing favorably to the protein-protein binding interface (at least according to Rosetta). Residue pairs with a highly positive scores are potentially harming the protein-protein interface.

For a detailed description of what each energy term means, please see Alford et al. https://doi.org/10.1021/acs.jctc.7b00125 Briefly, here are the meanings of the non-zero terms you're likely to see with `ref2015` scoring for residue-residue interactions between chains:

- fa_atr - the attractive component of the Lennard Jones (van der Waals) interaction
- fa_rep - the repulsive component of the the Lennard Jones (van der Waals) interaction
- fa_elec - the Coulombic electrostatic interaction
- fa_sol & lk_ball_wtd- Implicit solvation burial terms (from Lazaridis & Karplus, with modifications).
- hbond_sr_bb - "short range" (primarily alpha helical) backbone-backbone hydrogen bonds.
- hbond_lr_bb - "long range" (beta sheets & loops) backbone-backbone hydrogen bonds
- hbond_bb_sc - sidechain-backbone hydrogen bonds
- hbond_sc - sidechain-sidechain hydrogen bonds

(Other terms are either residue-internal energies, or will otherwise not show up for inter-chain residue-residue interactions.)

Take a moment to look at the residues pairs which are highly scored (either negatively or positively):

- According to the scoring, what sort of interactions are contributing to that score value?
- Looking at the docked structure in ChimeraX and visualizing those residues, does that interaction make sense?
- Can you determine which interactions are sidechain-based and which are backbone based?
- Would this analysis be helpful in determining potential mutational effect experiments?

Note that this analysis is somewhat basic. There are other protocols (e.g. the InterfaceAnalyzer) which are more dedicated to analyzing and quantifying protein-protein interfaces. Additionally, there are protocols which more directly probe possible mutational analysis experiments.

Advanced docking approaches.

The protocol that we used to dock the two structures is a general protocol to do protein-protein docking, but other specialized protocols exist.

SnugDock

Snugdock (https://doi.org/10.1371/journal.pcbi.1000644) is an antibody-specific docking protocol. In addition to the rigid body rotation and translation sampling, Snugdock optimizes the interchain heavy/light chain orientation, which has been shown to be important for antibody/antigen binding. Snugdock is available as a ROSIE server (https://rosie.rosettacommons.org/snug_dock).

AlphaFold & other ML structure prediction packages

Most machine-learning based structure prediction programs these days can be used to predict the structures of protein complexes. You simply have to give them the multiple chains you want to predict as a complex.

If you have time, try inputing the sequences of the A, B, H & L chains to a machine learning approach such as AlphaFold3 (https://alphafoldserver.com) or ColabFold (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). (For ColabFold, enter the sequence of each chain with a colon between it.)

Note that for protein-protein docking, in addition to the standard quality metrics like pTM and pLDDT, the pAE (predicted aligned error) metric is of particular importance. pAE measures the estimated error of residue-pairwise structure. As such, it's good to look at the between-chain pAE values, as that's the best estimate of how well the chains are placed (docked) relative to each other. If the ML model was able to successfully predict the monomer and thus the majority of a residues neighbors, that residue might still have a reasonably decent pLDDT value, despite not being in the proper orientation with respect to the binding partner. The pAE value of that residue with respect to the other residues in the chain should give a better sense of how accurately that residue was docked.

It's worth noting that most machine learning structure prediction packages still have difficulties with antibody/antigen complexes, particularly ones which don't have related complexes in the PDB. Extra scrutiny of such complexes is warranted.

It's also worth noting that ML complex predictions can be used as input to classical physics based docking approaches, if you want to further diversify and refine them.

AlphaRed

Alpha Red (https://www.biorxiv.org/content/10.1101/2023.07.28.551063v1 is a method which builds on AlphaFold multimer prediction, and combines it with a physics based replica exchange approach to further diversify and refine the docked structures. It has been shown to successfully find the native complex structures in cases where AlphaFold by itself does not. In particular, in a benchmark of antibody/antigen complexes, it improved the successful docking rate from 19% to 51%. AlphaRed is available as a ROSIE server https://r2.graylab.jhu.edu/apps/submit/alpha-red