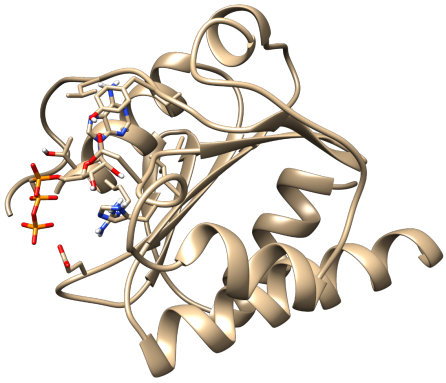
End-Semester Project Report 2024

Aim of the Project: To develop a model to predict the changes in holo-structures interacting with small molecule ligands using pre-existing crystal structures of apo and holo-proteins.

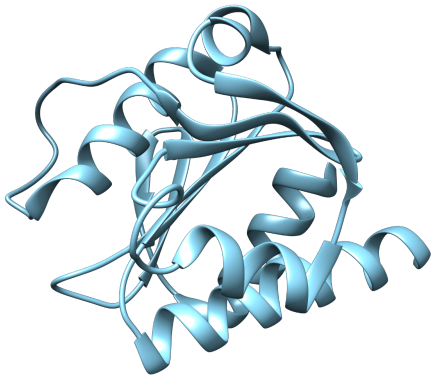
Abstract: Proteins when bound to small molecule ligands undergo certain conformational changes, especially in regions around the ligand binding site.This could lead to a loss/gain in function or efficiency of the protein. We want to study proteins in their bound and unbound states and see the changes that occur in RMSD between the two. Using this data we can model the changes that occur and use Molecular Dynamic simulations to see if we can try and predict these changes. If our model is successful in providing us with predictive binding sites and their size this can be used for Drug design and potentially to find a cure for Mycobacterium Tuberculosis.

Apo and Holo Proteins:

For the following study we use proteins for which we have both apo and holo forms. Apo proteins are those that don't have a ligand bound whereas Holo proteins have a ligand bound to them.



Holo protein



Apo protein

Dataset Filtration: We know that proteins having the same Uniprot IDs have the same sequence. So we look for Uniprot IDs with two proteins one

of which has a ligand and one of which doesn't. We filter out NMR structures since those are dynamic and don't have a fixed pose. Also among the remaining, structures with resolution less than 3.5Å are filtered out since those are too inaccurate to analyse small changes of that order. We also remove proteins with ligands under 100 Da for similar reasons and those over 1000 Da keeping in mind Lipinski's rule of 5. After these steps we get 798 pairs of apo-holo structures.

Methods used: To analyse what changes happen in the protein due to the binding of a small molecule we first analyse the Root Mean Square Deviation (R.M.S.D) between the apo and holo protein by first aligning them using the chains that are common between the two (using 3D Least Square Algorithm). The R.M.S.D thus obtained is the Global R.M.S.D. Now to analyse the changes in the immediate surroundings of the ligand we chop out a binding pocket of 5Å from the holo protein (since this is range of Van der Waals interactions) and align only that specific region to the same residues on the apo protein and find the R.M.S.D of just this region (local R.M.S.D). This will help us pinpoint what the changes due to the ligand interactions are.

3D Least Square Algorithm:

We input lists of coordinates with atom-atom correspondences, in this case the residues from the common Uniprot IDs atom wise will be aligned. Only if a certain atom of a residue is present in both the files do we add the coordinates to an array. Only atoms from chains that are corresponding to Uniprot ID are used to populate the array in cases where a protein has multiple chains. The two arrays generated from the apo and holo file are then fitted by finding a translation vector T and a rotation vector R which minimises the R.M.S.D between corresponding atoms. This calculated R and T are then applied to all the atom coordinates in the PDB file to see the superimposed proteins. It requires atom atom correspondences which in our case are all the common atoms other than Hydrogen.

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - (\mathbf{Y}_i\mathbf{R} + \mathbf{T})\|^2}$$

Local and Global:

To analyse the global changes that happen in the protein structure we find the chain in the protein that has the Uniprot ID we have used to match the structures. We do this using a line that starts with DBREF and see which of the lines starting with this has the corresponding Uniprot ID. We now match the coordinates from this chain in apo protein to the coordinates from holo protein of the same chain and only retain the common residues and atoms from the two. We then align them using the 3D square fit algorithm as described by Kearsley (Kearsley 1989).

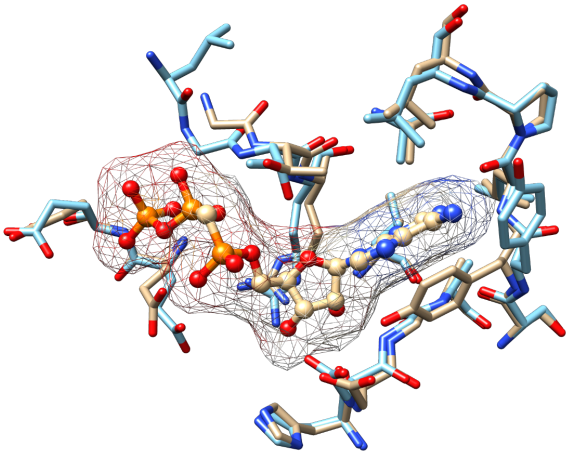
To analyse the local change that happens specifically due to binding of the ligand we chop out of 5 Å binding pocket from the holo protein. If any atom of a nearby residue is within 5 Å of any atom of the ligand we add that residue to the chopped PDB. Now if these residues all belong to the same chain that was initially used to align them, then we align the residues in the chopped binding pocket to those residue numbers in the apo protein and get an RMSD from this. We initially only focussed on single chain pockets because in case of multi chain pockets one of the chains may not be a part of the common Uniprot and this could cause problems when aligning. This is the local RMSD induced by the ligand binding. The local residues can be analysed in the binding pocket and prediction of residue environments for certain ligands can be done from this.

- Apo protein (PDB ID - 6GYY)

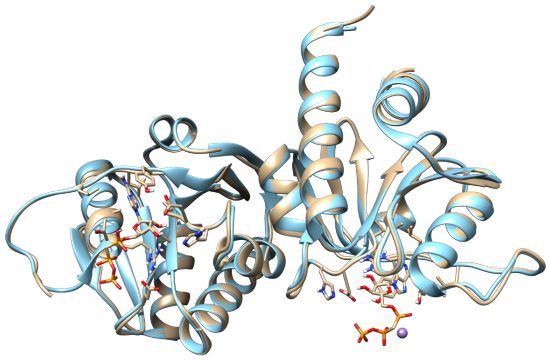
- Holo protein (PDB ID - 6GYX)

The structure with the mesh is the ligand in this case is APC.

DIPHOSPHOMETHYLPHOSPHONIC ACID ADENOSYL ESTER



Local Alignment



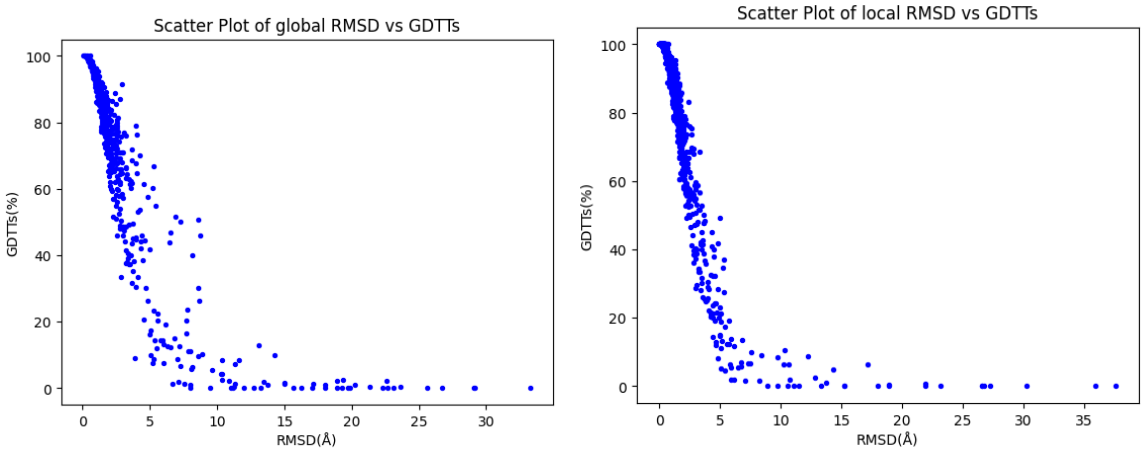
Global Alignment

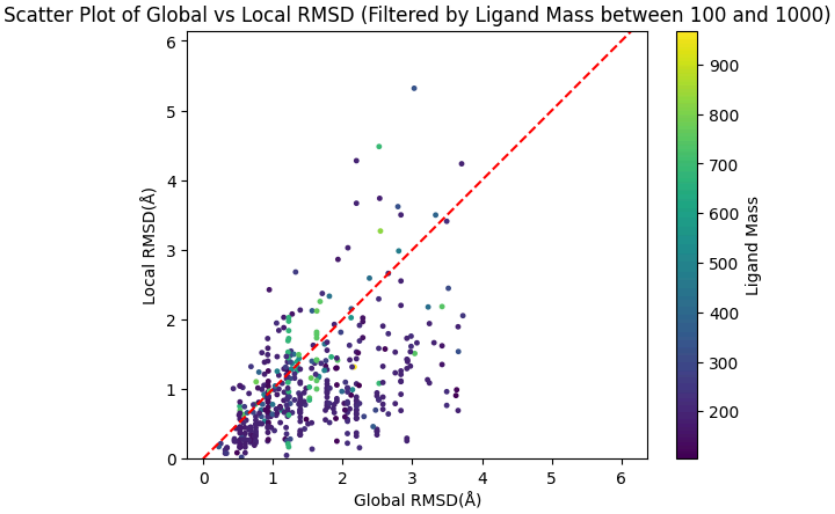
Structural Overlap and GDTT score:

The 3D least square fit algorithm also calculates the structural overlap which is the number of atoms overlapped, given a distance cutoff. If m out of n atoms are within the SODC (structural overlap distance cutoff) then the overlap is m/n*100. We calculate this Structural overlap at 3 different distance cutoffs i.e. 1Å, 2Å, 3.5Å and calculate a GDTT (Global distance test total) score which is essentially going to be an average of these.

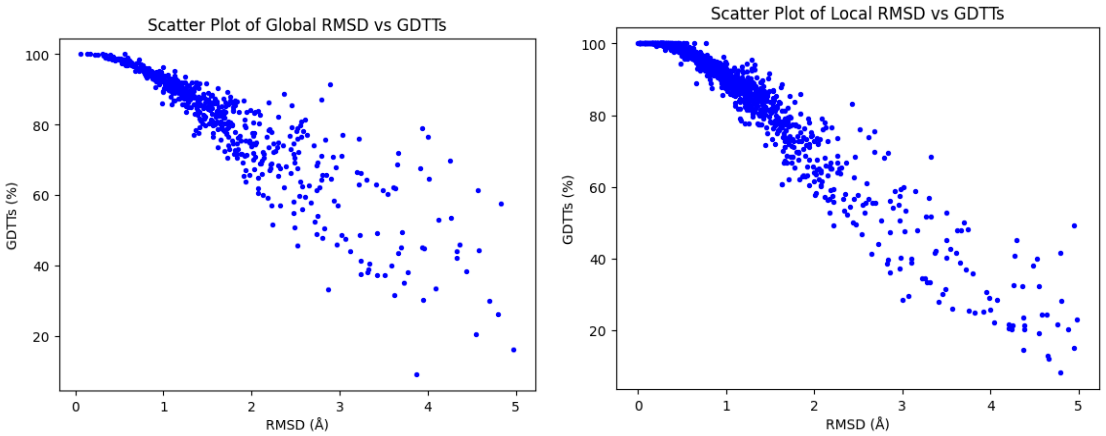
$$\mathrm{GDTTS} = \frac{1}{3} \sum_{i \in \{1\text{Å},2\text{Å},3.5\text{Å}\}} \mathrm{SO}_i$$

Results:

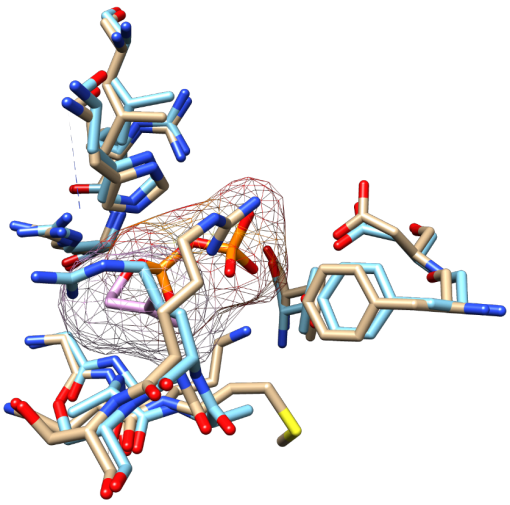




This is what the data for Local RMSD vs GDTTs and Global RMSD vs GDTTs looks like. This makes sense because as more atoms are out of the distance cutoff, RMSD would increase and conversely structural overlap would decrease so the first 2 scatter plots make sense. However the reason for extremely high RMSD in some of the global cases and correspondingly in the local cases is due to proteins with the same sequence having different orientations in space. This causes them to have a high RMSD. Below are plots where global and local RMSD are bounded below 5Å.



Now to see what cases have local RMSD more than Global we make the second plot which gives a metric of how much of the change is local compared to overall difference in the 2 proteins. The red dotted line is y=x and we can see there are numerous points above the lines which show that in some cases the local RMSD is higher than the global which shows a significant part of the change is probably due to ligand binding. Following is one of the cases:



Ligand in this case is IPE. 3-METHYLBUT-3-ENYL TRIHYDROGEN DIPHOSPHATE
Holo PDB id - 7S0M
Apo PDB id - 7S0H

Future work:

- We can use the data we have to check if certain types of residues have predictable behaviour, for instance we may find hydrophobic residues don't tend to move much.
- We can use it to analyse ligand specific behaviour i.e. comparing ligand binding pockets and seeing if they have a recurring pattern in type of binding pocket.
- Training Artificial Intelligence or Machine learning models to predict binding pockets in proteins that can be targeted for drug design.
- Run MD simulations to see if any of the known binding sites can be predicted using our model and hence validate it.

References:

1) On the orthogonal transformation used for structural comparisons. Simon K Kearsley 1989, Acta Cryst 1989, A45, Pg 208-210.