Evianne Rovers
M. Sc. Student at Structural Genomics Consortium (SGC) Toronto/Pharmacology and Toxicology Department, University of Toronto

Distinguishing between catalytic and non-catalytic pockets in the ligandable human genome

My goal is to find druggable pockets in human enzymes that are non-catalytic. These non-catalytic druggable pockets may then be exploited for proximity pharmacology (ProxPharm), a novel paradigm in drug discovery where chimeric compounds bring two proteins in close proximity to elicit an effect of one protein on the other[1]. For example, PROTACs simultaneously bind an E3 ligase and a substrate protein, leading to ubiquitination and subsequent degradation of the substrate[2]. ProxPharm compounds should not inhibit the recruited enzyme (Figure 1). Therefore, a binding location which is not located in the catalytic domain, or is in the catalytic domain but distant from the catalytic site, is favored.

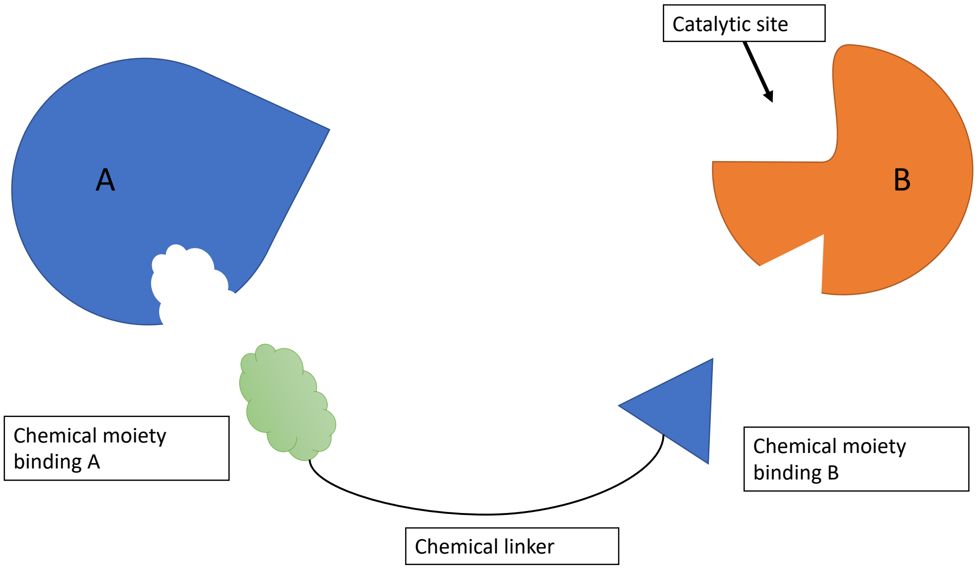


Figure 1. *ProxPharm compound binds enzyme (B) and substrate (A) without inhibiting the enzyme.*

In our lab, Setayesh Yazdani initiated the Ligandable Genome project based on the project of Jiayan Wang et al.[3] Jiayan's project analysed exclusively human protein structures with bound ligands and assessed the druggability of the corresponding protein domains that were occupied by at least one drug-like ligand in the PDB. Setayesh extended the analysis by mapping cavities in all human proteins in the PDB, whether or not they were bound to small-molecule ligands.

She identified the pockets and calculated their properties by using the icmPocketFinder method in ICM (Molsoft, San Diego).

The next important step in this project is to distinguish between catalytic and non-catalytic pockets found by icmPocketfinder in human enzymes. To do this I measure the distance between catalytic/active site residues available from the Mechanism and Catalytic Site Atlas (M-CSA) database[4] or available from the UniprotKB database[5] and pockets found with IcmPocketFinder (Figure 2). One of the benefits from this approach is that the proximity of the pocket to the catalytic domain is quantified. Figure 3. shows that the minimum distance between the catalytic domain of 5'-nucleotidase (CD73) (PDB code 6TVE[6]) and pocket 13 is 13 Å.

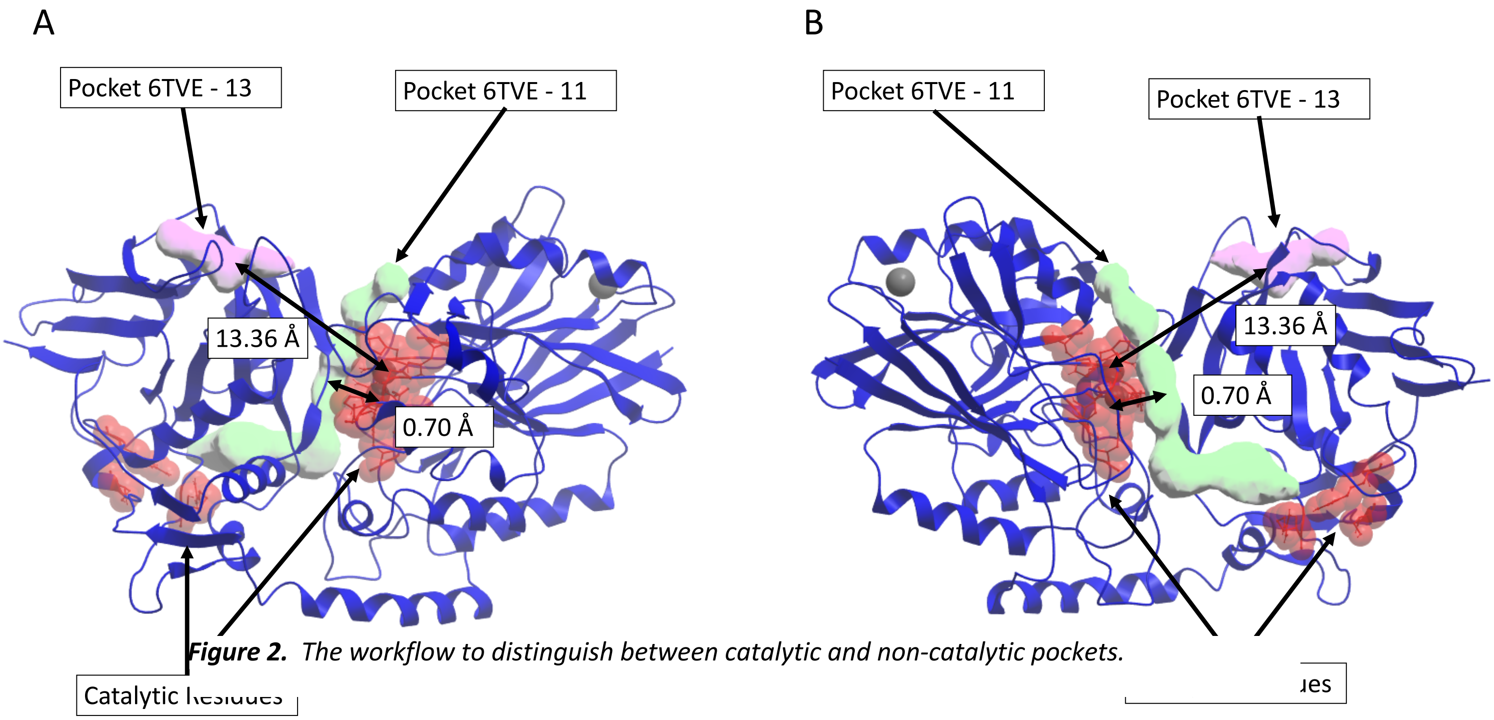
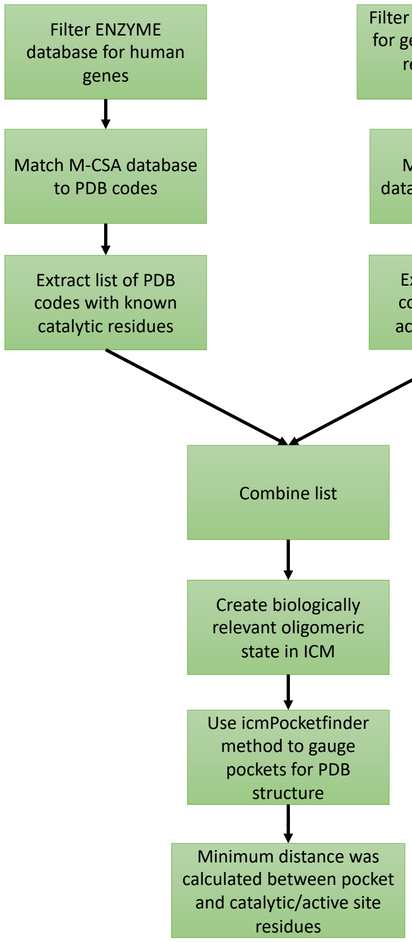


***Figure 2.** The workflow to distinguish between catalytic and non-catalytic pockets.*

***Figure 3.** Discriminating between catalytic and non-catalytic pockets for 5'-nucleotidase (CD73) (PDB code 6TVE)[6]. Chimeric compounds targeting the purple pocket to recruit the enzyme to a new substrate are far from catalytic residues, and therefore unlikely to inhibit the enzyme. A) Pocket 13 (purple) distance to catalytic residues is 13.36 Å. B) Pocket 11 (green) distance to catalytic residues is 0.70 Å and distinguished as catalytic site.*

Methods:


Filter ENZYME database for human genes
Match M-CSA database to PDB codes
Extract list of PDB codes with known catalytic residues
Filter UniprotKD database for genes with active site residues noted in UniprotKD
Match UniprotKD database to PDB codes
Extract list of PDB codes with known active site residues
Combine list
Create biologically relevant oligomeric state in ICM
Use icmPocketfinder method to gauge pockets for PDB structure
Minimum distance was calculated between pocket and catalytic/active site residues

Step 1: Compile_PDB_db

1. UniprotID's and gene names were extracted from the Expasy ENZYME database and filtered for human genes. UniprotID's and PDB codes were obtained from the UniProt database. Both tables were joined by UniprotID and the joined table was called enzyme_GN_UID_PDB. The table was filtered for human gene only.
2. M-CSA database was uploaded in Molsoft ICM-Pro and the catalytic residues were linked to corresponding UniprotID in the human_enzymes_all table. PDB codes for enzymes without catalytic residues specified in M-CSA were filtered out of the humenz_MCSA_db table.
3. Table was generated on the UniprotKB website for human genes and column added for active site. Table was downloaded in excel format and for the human genes without active site known were filtered out. The excel sheet was uploaded in Molsoft ICM-Pro. Table was called humenz_UniprotKD_db.

4. humenz_MCSA_db and humenz_UniprotKD_db were both joined together to analyse for duplicates and check for congruities in the information about active site residues. The catalytic residues obtained from the M-CSA database are noted in the 'M CSA db' column and residues numbers obtained from the UniprotKD database were noted in the 'UniprotKD db' column. The final list was called hum_enz_all. The table was filtered for unique PDB codes in the table humenz_unique_PDB and was also filtered for unique gene names in the table humenz_unique_gene.

Step 2: Convert_objects

5. Each PDB file in the humenz_unique_PDB table was uploaded in ICM. Ligand and water molecules were deleted and peptides less than 15 amino acids were removed from the object. Afterwards, if the object contained less than 2000 amino acids, it was converted to an ICM object and the biologically relevant oligomeric state was generated. PDB structures that could not be converted to ICM objects were flagged in the 'Object flag' column of humenz_unique_PDB table.

Step 3: icmPocketfinder

6. For each ICM object, the icmPocketfinder method was run against each isolated peptide chain using the default settings. The icmPocketfinder generated a table with information about the volume, hydrophobicity, buriedness and area of the pockets, along with 3D objects of the pockets. The pocket table was saved and the 3D objects of the pockets were saved individually. ICM objects that did not have pockets were flagged in the 'Pocket flag' column of humenz_unique_PDB table.

Step 4: Pocket_Analysis *(Distance calculations between catalytic residues and pocket)*

7. The ICM object, related pockets and pocket table were opened.
8. To ensure proper numbering of amino acids in our protein structures, residues from the ICM object were renumbered by aligning the individual peptide chains to the Uniprot reference sequences found in "UP000005640_9606.fasta". Each peptide chain was aligned separately, because a PDB structure could contain multiple peptide chains of different genes.
    a. ICM objects that contained one (or more) peptides of the TITIN_HUMAN gene were analysed using the 'Step_4_Adjusted_TITIN_HUMAN script. These structures were renumbered incorrect if the sequence contained residues above residue number 32757. If an ICM object contained a TITIN_HUMAN peptide with residues above 32757, these peptides were not renumbered and flagged in the humenz_unique_PDB table in the column 'TITIN_HUMAN flag'.
9. The UniprotID, percentage of sequence similarity, and beginning and end residue number of the peptide chain were recorded in the pocket table for each pocket.
10. Then, the residues in the ICM object with a 2.8 Å distance from the pocket were taken as residues lining the pocket. These pocket residues are located near the surface of the pocket and together form a pseudo "sequence" for each pocket.
11. After obtaining the pocket residues, the beginning and ending residue numbers of the pocket were identified and noted in the table. The pocket length, denoted as the length between the beginning and ending residues, was calculated. Also, in the pocket table the peptide chain, in which the pocket was located, was noted.

12. Large gaps at the beginning or end of the pocket residue sequence (either between the first and second residue or last but one and last residue) were flagged. The cut-off was >50 residues in between the two residues. If flagged, the beginning or ending residue was deleted.
13. Using the sequence function, and the start and end residue number, the pocket sequence was obtained.
14. For cases that have valid length, but were (partly) in a non-human chain or N-term tag (which can be identified by residue number 0 after renumbering of the structures), they were denoted to have no valid length.
15. In the pocket table, the catalytic residues from the M-CSA database (table M_CSA_db) were added based on the UniprotID obtained during the sequence aligning of the individual peptide chains. This was done to ensure that in PDB structures with multiple protein domains of different genes; only the catalytic residues relevant to the peptide chain were taken into consideration.
16. If catalytic residues for the UniprotID were not available in the M-CSA database; the active site residues noted in the UniprotKB database (table UniprotKD_db_act_site_sheet1) were added to the pocket table according to UniprotID.
17. Then, the minimum distance was calculated between the pocket and catalytic residues present in the peptide chain.
18. Catalytic residues of an enzyme annotated in the MCS-A or UniprotKB database may be missing from a structure of the enzyme available from the PDB for two possible reasons: (1) the protein domain available in the PDB is not the catalytic domain. In this case, any druggable pocket found in the structure could in principle be exploited by ProxPharm compounds without affecting the catalytic activity of the enzyme. (2) The catalytic residues are in a disordered region of the structure. If the case, they may be next to identified pockets, and these structures are filtered-out in later analysis.
19. Next, there might be gaps/disordered regions in the structures. Therefore, the distance between the flanking residues of these gaps in the structure and pockets was calculated. The missing residues numbers were obtained by aligning the sequence of the peptide chain to the Uniprot reference sequence and identifying the missing residues and the corresponding flanking residues.
20. Lastly, the peptide chains could include an expression tag at either the N- or C-terminus of the sequence. These residues were obtained by aligning the peptide chain sequence with the Uniprot reference sequence and identifying the residues present in the structure before the first aligned residue (PDB start) and after the last aligned residue (PDB end). Then distance between these expression tag residues was calculated.

Results:

- 1487 human enzymes could be identified with known catalytic residues. (Excel table humenz_unique_gene)
- That corresponds to 19510 PDB structures. (Excel table humenz_unique_PDB)
- 18,654 PDB structures could be converted to ICM biomolecule objects. (Excel table humenz_unique_PDB (without “+” in Object Flag column))

- 140,276 pockets were found for 18,498 ICM objects. (Excel tables: humenz_unique_PDB (without "+" in Object flag and Pocket Flag column) and POCKETS_hum_enz_all)
- Distance calculations showed catalytic and non-catalytic pockets.
- All scripts and data are in the icm-file '20201125_Ligandable_Human_Genome_project.icb' that can be open with the free ICMBrowser from Molsoft.com.

Next steps:

- Determine which pockets are druggable.
- Removing duplicates (same pocket represented by multiple PDB codes).

References:

1. Ribeiro, A. J. M. *et al.* Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).
2. Yazdani, S. & Schapira, M. A Gentle Introduction to The Ligandable Genome Project Method 1. (2020). doi:10.5281/ZENODO.3677177