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: InterPro analysis

As describe in my previous post, my goal is to discover non-catalytic druggable pockets in human enzymes. These pockets could potentially be exploited for the design of ProxPharm compounds (chimeric compounds that bring two proteins in close proximity to elicit an effect of one protein on the other[1]). An essential aspect for the design of ProxPharm compounds is that the chemical moiety binding the enzyme binds to a non-catalytic pocket, because the compound should not inhibit the activity of the recruited enzyme.

Previously, the methods of both Jiayan Wang's[2] and Setayesh Yazdani's[3] project were used to identify the pockets in all human proteins available in PDB regardless if they were bound to small-molecule ligands. For this, the icmPocketFinder module was used in ICM software (Molsoft, San Diego). In my previous post, I categorized pockets as either catalytic or non-catalytic by measuring the distance between the pocket and catalytic residues present in the structure. The catalytic residues were obtained from the Mechanism and Catalytic Site Atlas (M-CSA) database[4] or the UniprotKB databse[5] (Figure 1 (Green))[6].

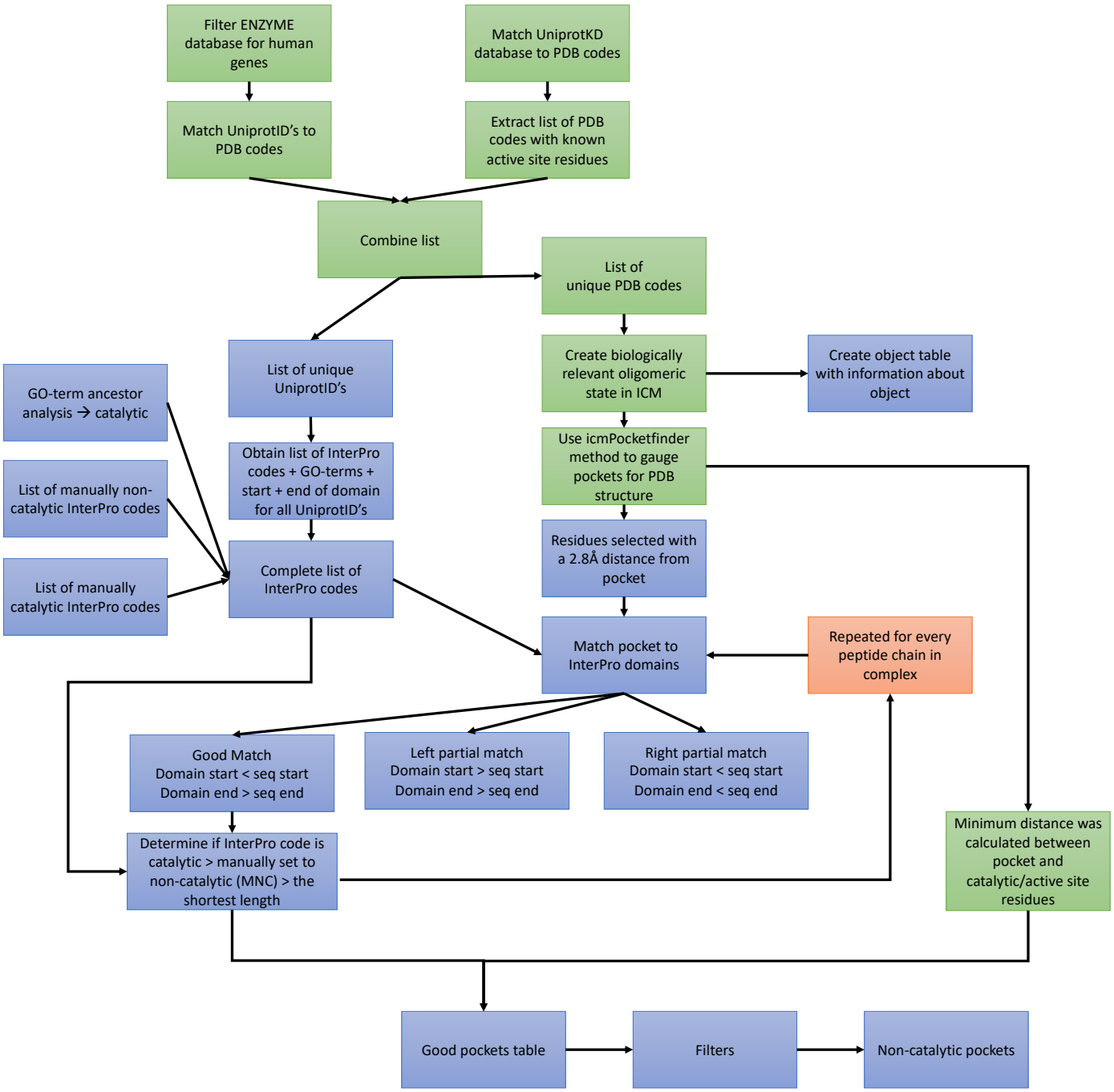

Figure 1. *The workflow to distinguish between catalytic and non-catalytic pockets. Green boxes represent first approach (first post) and blue boxes extended second approach.*

This method yielded a low amount (1478) of human enzymes to be analyzed, because both databases contained limited information on catalytic residues for human enzymes. Therefore, a second approach was tested (Figure 1 (Blue)), namely identifying whether a pocket is in a catalytic domain by matching the domain information from the InterPro database[7] to the pocket (Figure 2).

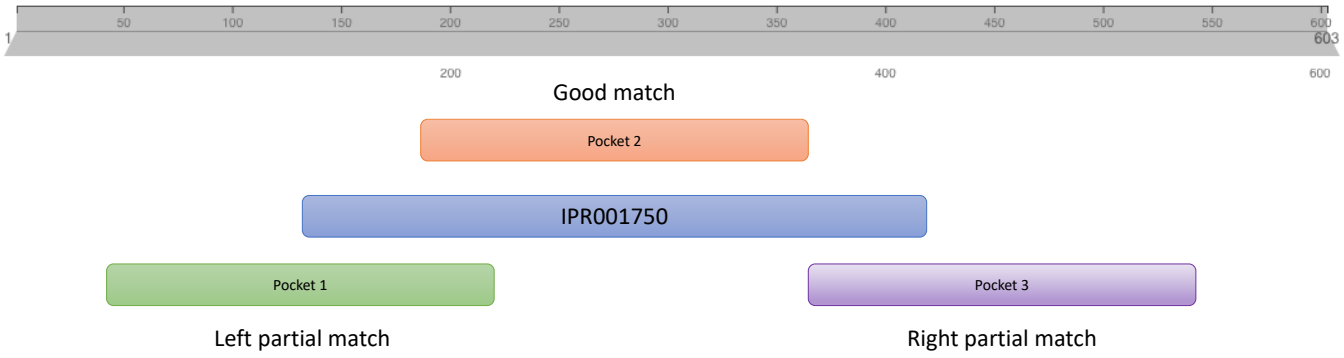

Figure 2. *Example of InterPro code with 3 pockets, pocket 1 is let partial match, pocket 2 is good match and pocket 3 is right partial match. (IPR001750: NADH:quinone oxidoreductase/Mrp antiporter, membrane subunit[7])*

Methods:

Step 1a: Compile_PDB_db

1. UniprotID's and gene names were extracted from the Expasy ENZYME database and filtered for human genes (table enzyme_in_genename). UniprotID's and PDB codes were obtained from the UniProt database (table enzyme_in_PDB). Both tables were joined by UniprotID and the joined table was called enzyme_GN_UID_PDB.
2. Table was generated on the UniprotKB website for human genes and column added for EC number. The corresponding PDB code for the UniprotID was obtained from the table enzymes_in_PDB. The table was called enzyme_UKB.
3. Both tables (enzyme_GN_UID_PDB and enzyme_UKB) were joined together to form the table Humenz_all and filtered for either being in the ENZYME database or known EC class number. The new table was called enzyme_proteins.
4. The unique UniprotID's were stored in the table UniprotID and the unique PDB codes are recorded in the table PDB.

Step1b: Python scripts for obtaining table go_interpro and M_CSA_db.

5. The table UniprotID is converted to excel file. Then in python, the excel file is opened and for each UniprotID, the website: "http://www.ebi.ac.uk/interpro/api/entry/all/protein/reviewed/"+Uniprot[x]+"/" is opened. The information is stored in json format and the accession, source database, GO terms, UniprotID and start and end residue are recorded in a .tsv file.
6. Then, the InterPro codes with the accession 'interpro' were recorded in the table named 'go_interpro'.
7. Next, a list of unique GO-terms is formed. The ":" signed is replaced for %3A and the list is converted to excel file. The excel file is opened for each UniprotID and the website "https://www.ebi.ac.uk/QuickGO/services/ontology/go/terms/"+[go-term]+/ancestors?relations=is_a%2Cpart_of%2Coccurs_in%2Cregulates" is opened. The information is stored in json format and the GO-term and its ancestors are written into a .tsv file (table go_terms_ancestor).

8. All the InterPro codes of which the GO-terms had the ‘catalytic activity’ ancestor (GO:0003824) were flagged as Catalytic in the go_interpro table.
9. A list was created with known non-catalytic InterPro codes (manually non catalytic table (MNC)). These InterPro codes were flagged in the MNC column in the go_interpro table.
10. The excel file with UniprotIDs is used to obtained the catalytic residues in the M-CSA database. In python, the excel file is opened and for each UniprotID, the website: "https://www.ebi.ac.uk/thornton-srv/m-csa/api/entries/?format=json&entries.proteins.sequences.uniprot_ids="+Uniprot[x] is opened. The information is stored in json format and UniprotID and catalytic residues are recorded in a .tsv file (table M_CSA_db).

Step 2: Convert_objects

11. Each PDB file in the 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 the PDB table.

Step 3: icmPocketfinder

12. For each ICM object, the icmPocketfinder method was run against each converted object 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 the PDB table.

Step 4: Object_info

13. The ICM object was opened.
14. 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 not renumbered, because these structures were renumbered incorrect if the sequence contained residues above residue number 32757. These objects were flagged in the PDB table in the column ‘TITIN_HUMAN flag’.
15. The UniprotID, beginning and end residue number of the peptide chain were recorded in the ‘Object table’ for each peptide chain as well as the pocket molecule name.
16. Next, there might be gaps/disordered regions in the structures. The missing residues numbers were obtained by aligning the sequence of the peptide chain to the UniprotID reference sequence and identifying the missing residues and the corresponding flanking residues. These were recorded in the Object table for each peptide chain.
17. 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 UniprotID 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). These residue numbers for these expression tag residues were recorded in the Object table.

18. In the Object 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.
19. The active site residues noted in the UniprotKB database (table UniprotKD_db_act_site_sheet1) were added to the Object table according to UniprotID.
20. Catalytic residues of an enzyme annotated in the M-CSA 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.
    a. So, the catalytic residues were analyzed for the presence in the object. If present in a region of missing residues, it was flagged in missing_cat_res_flag column. If not present and outside of the residue range in the object, it was flagged in the outside_chain_flag column.
21. Afterwards, the distance between peptide chains and the residues of another peptide chains was calculated and if < 5 Å, the distance and the peptide chain were recorded in the Distance2 and Protein2 columns of the object table. These proteins are considered inhibiting proteins as they bind to the catalytic site of the other protein. In the pocket analysis, the pockets in these inhibiting proteins as well as the pockets on the surface between the inhibiting protein and other protein will be filtered out.
22. When no catalytic residues are available from the MCSA_db or UniProt_KB db then the peptide chains were matched to the InterPro database. For every catalytic domain that was present in a peptide chain, the distance was calculated to other peptide chains. If < 5 Å, the distance and the peptide chain were recorded in the Distance2 and Protein2 columns of the object table and considered as an inhibiting protein.

Step 5: Pocket_Analysis

23. The renumbered ICM object, related pockets, pocket table and object table were opened.
24. 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. For example, the pocket could be on the surface between two homodimers and thereby have a sequence in both peptide chains in the object.
25. For each pocket, a table was generated containing the residues and in which peptide chain the residues were located.
26. If the pocket was located in multiple peptide chains, the following steps were repeated for every peptide chain.

a. 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.
b. If the residue sequence was longer than 6 residues it was denoted to have valid length. Pockets with less than 6 residues are shallow pockets and would have less residues to have interactions with.
c. Then based on the UniprotID obtained in step 8, the corresponding list of InterPro codes were retrieved from the go_interpro table. For every InterPro code, the shared residues percentage was determined by calculating the amount of pocket residues that were in the range of the domain divided by the total amount of pocket residues.
d. In the case that domain start (dom_start) is smaller than pocket sequence start (seq_start) and domain end (dom_end) bigger than sequence end (seq_end), it was considered a good match and it was recorded in the g_match_result_2 array with a prefix of the peptide chain. After the analysis for all peptide chains, all the good matches for all peptide chains were recorded in the ip_good_match column in the pocket table.
e. In the case that dom_start was bigger than seq_start and seq_end smaller than dom_end, it was called a right partial match. If the dom_start was smaller than seq_start and seq_end was bigger than dom_end, it was called a left partial match. (Figure 2)
f. For either partial match, the number of matching residues (p_match_percent) for the domain was calculated by the number of matching residues divided by total amount of domain residues. The number of matching residues for the pocket sequence (seq_match_percent) was calculated by the number of matching residues divided by total amount of pocket residues.
g. If the both the p_match_percent and seq_match_percent were above 0.90, the match was recorded as good match in the g_match_result_2 array with a prefix of the peptide chain. Otherwise, it was recorded in p_match_result_2 array with a prefix of the peptide chain with the percentage of domain match (p_match_percent).
h. Afterwards, all the good matches were grouped in g_m_ipr table. Good matches with 0 % shared residues were filtered out.
i. Of the domains that overlapped the same region and were 50 residues longer than the smaller domain, the smaller domain was retained and the larger domain was filtered out.
j. Next, the g_m_ipr table was checked for catalytic and/or manually non-catalytic (MNC) domains. If present, the domains that were not flagged as catalytic or MNC were filtered out. This was to eliminate domains that overlap the same region, but were flagged differently. This ensured that only the domains that had known information about being catalytic or being non-catalytic were retained for further analysis. If none of the domains were flagged catalytic or MNC, this step was skipped.
k. Afterwards, if the domains overlapped the same regions, the smaller domain was retained and the larger domain was filtered out.

l. Lastly, the table was sorted on percentage of shared residues and the InterPro code with the highest percentage was recorded as the IPR id in the column iprid with peptide chain prefix along with the domain name, length and go id. Also, whether it was flagged as catalytic or MNC.

27. Then, the minimum distance was calculated between the pocket and catalytic residues present in the peptide chain.
28. Next, the distance between the flanking residues of missing residues in the structure and pockets was calculated.
29. Also, the distance between the expression tag residues of the peptide chain and the pocket was calculated.
30. Pockets that are in an inhibiting protein or on the interface between a protein and inhibiting protein were flagged in the Obstruct_flag column of the pocket table. An inhibiting protein is defined as a protein within a < 5 Å distance from the catalytic residues/domain of another protein in the complex as calculated in step 20 and 21.
31. All the pockets that were on the interface between two or more proteins were flagged in the Interface_flag column in the pocket table.

Step 6: Combine Pocket tables

32. All the pocket tables were combined to make one table called good_pockets.

Step 7: Filtering non catalytic pockets

33. Pockets that were in a non-catalytic domain of the protein or in the catalytic domain, but have a >7 Å distance from the catalytic residues were transferred to the non_cat_pockets table.
34. When none of the domains of a given protein are annotated as "catalytic" in the InterPro database and catalytic residues are not provided in the M-CSA or UniprotKB databases it is impossible to know whether a pocket identified in this protein is catalytic or not. In this case, all pockets of that protein are ignored.
35. Furthermore, only pockets that were >5 Å distance from the missing residues were retained in the non_cat_pockets_2 table.
36. Also, pockets that were marked to be present in an inhibiting protein or on the interface of the inhibiting protein and the other protein (Obstruct_flag column) (as calculated in step 29) were filtered out.
37. Lastly, pockets that did not have valid length (< 6 residues from a 2.8 Å distance from the pocket; calculated in step 25b) were filtered out.

Results:

- **178,022** pockets were identified in **23,621** structures representing **2310** proteins.
- **71,592** non-catalytic pockets in **11,367** structures representing **1824** proteins.

Next steps:

- Remove duplicate pockets when different PDB codes are available but for a single protein.
- Filter pockets for druggability (volume, area, hydrophobicity and buriedness)

References:

1. Gerry, C. J. & Schreiber, S. L. Unifying principles of bifunctional, proximity-inducing small molecules. *Nat. Chem. Biol.* **16**, 369–378 (2020).

2. Wang, J., Yazdani, S., Han, A. & Schapira, M. Structure-based view of the druggable genome. *Drug Discov. Today* **25**, 561–567 (2020).
3. Yazdani, S. & Schapira, M. A Gentle Introduction to The Ligandable Genome Project Method 1. (2020) doi:10.5281/ZENODO.3677177.
4. 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).
5. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
6. Rovers, E. & Schapira, M. Distinguishing between catalytic and non-catalytic pockets in the ligandable human genome. (2020) doi:10.5281/ZENODO.4294099.
7. Hunter, S. *et al.* InterPro: The integrative protein signature database. *Nucleic Acids Res.* **37**, 211–215 (2009).