Charles University

Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Adam Král

Framework for retrieval and analysis of proteins apo and holo forms from PDB
Framework pro získání a analýzu apo- a holo- form proteinů z PDB

Bachelor's thesis

Supervisor: doc. RNDr. David Hoksza, Ph.D.

Prague, 2022

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Acknowledgement

I would like to thank my supervisor, doc. RNDr. David Hoksza, Ph.D., for the patience he had with me, for his advice, and for his positive outlook.

Abstract

We developed a software framework that allows the analysis of ligand-free (apo) and ligand-bound (holo) forms of proteins that are accessible in PDB. The software downloads the current version of the PDB, divides the structures into groups of the same molecules, and these into apo and holo forms. Finally, it is possible to analyze pairs of apo and holo structures with respect to their different structural characteristics. In addition to the software work itself, we present the results of selected analyses of the current version of the data in the PDB. We also verify the results against previous work.

Abstrakt

Vyvinuli jsme softwarový framework, který umožňuje analyzovat a porovnávat apo (bez ligandu) a holo (s ligandem) strukturní formy proteinů přístupných v PDB. Software stáhne aktuální verzi PDB, rozdělí struktury do skupin stejných molekul a rozliší zda se jedná o apo či holo strukturní formu. Nakonec je možné analyzovat dvojice apo a holo struktur s ohledem na jejich odlišné strukturální charakteristiky. Kromě samotné softwarové práce prezentujeme výsledky vybraných analýz aktuální verze dat v PDB. Výsledky také ověřujeme oproti výchozímu výzkumu.

2 Table of contents

3 Motivation

An ever-growing database of experimentally resolved protein structures, Protein Data Bank (PDB, Berman et al., 2000)**,** allows studying protein structures in silico**.** It contains protein structures crystallized with and without ligands.

Ligand-protein interactions are part of the system of the living cell. Ligands may induce a change in protein structure, capable of altering the role or the function of the protein in the cell. Most of the drugs are ligands. Therefore, exploring and learning about differences between ligand-free (apo) and ligand-bound (holo) structures may be in the interest of researchers, or secondhandedly, their machine learning pipelines.

For example, in the task of ligand binding site prediction, identifying if and where the protein of interest could bind a ligand (any), we are presented with an apo structure. If we were to use a data intensive machine learning method (requiring a large training dataset), we would want to identify as many apo-holo pairs, for which their structure is resolved, as possible.

Or, for the same task, we might want to source a special dataset of apo-holo pairs, where the ligand binding site in apo structure is not as evident as in the holo structure, perhaps it is blocked by a mobile domain that dissociates in the event of ligand binding. An algorithm trained just on holo structures – on prediction effectively treating the apo structure as a holo structure minus the ligand – wouldn't detect the binding site. However, having the dataset of those so-called cryptic binding sites, the researcher could explore it, prototype a new algorithm, and use the dataset to train it and evaluate it.

We hereby present a software framework implementing the common functionality as required by the tasks above. Downloading the current version of the PDB, dividing the structures into groups of the same molecules, and then dividing these into apo and holo forms. Generating machine-readable dataset (JSON) in each step. Finally, it is possible to analyze the pairs of apo and holo structures with respect to their different structural characteristics. We include scripts to execute the pipeline on Czech National Grid Infrastructure (Metacentrum).

4 Proteins and ligands

Proteins are ubiquitous in living cells. They have various functions – for example some determine the shape of the cell by forming a cytoskeleton, others – enzymes – catalyze chemical reactions with metabolites, and some are parts of signaling pathways – biochemical cascades – that allow the cell to sense and react to its surroundings e.g. by changing the gene expression.

Ligands are comparatively smaller molecules (to distinguish it from the binding partner, e.g. a protein) that non-covalently, reversibly, bind to biomolecules, such as proteins. (In this work, we do not consider polynucleotides (DNA, RNA) as ligands to proteins.) They include potential drugs but also (natural) substrates to enzymes, signaling molecules, etc.

Binding a ligand may induce a conformational change (i.e. change in the protein structure). This may result in alteration of the protein function.

For example, the ligand latrunculin A binds actin, a cytoskeleton protein monomer, in such a way it prevents the polymerization in cytoskeleton filaments (Morton et al., 2000). (In theory this could be due to the ligand presence itself, sterically conflicting with the interface between the actin monomers, and not due to an alteration of conformation of the protein per se, however (Morton et al., 2000) shows on the resolved structure that the ligand indeed is deeper in the actin structure and it 'controls' the more distant subunit interface.)

Or, caffeine, a ligand to the A2A receptor, prevents a natural ligand, adenosine, from binding to the receptor (Snyder et al., 1981). Caffeine is in chemical structure similar to adenosine so it *perhaps* binds to the same site, then however, it is dissimilar enough to *not* cause *the* conformational change of the receptor like adenosine does, as it has been shown adenosine promotes sleep (Huang et al., 2011) while caffeine acts as a stimulant (Snyder et al., 1981).

4.1 Protein structure

Proteins first and foremost consist of one or more linear chains of amino-acid residues (one or more polypeptides). Proteins of multiple chains are also referred to as protein complexes.

The sequence of the polypeptide residues is called the *primary structure*. The peptide chain is synthesized in the living cell linearly, residue-by-residue adding on the nascent chain.

Protein *backbone* is the linear chain of atoms of amino acid residues minus their side-chains – atom groups specific to each amino acid attached to its alpha carbon (which is itself part of the backbone). The chain is somewhat flexible, as bonds around the alpha carbon generally allow more than one torsion angle.

Eventually after (or during) the synthesis the chain folds into a more stable conformation or *fold*.

The patterns of hydrogen bonds of the backbone atoms in the fold can be classified into different types of so-called secondary structure. Screw-like alpha helices and somewhat planar beta sheets are the most common types of secondary structure, strongly regular.

The overall chain's three-dimensional fold is called the *tertiary structure*.

4.2 Protein domains

Protein domains can be defined as parts of the tertiary structure that are independent, in the sense that the structure-stabilizing contacts between residues are primarily contacts between the residues within the domain. As a result, they are somehow rigid, but may undergo movements wrt. each other.

4.3 Determining protein structure

Protein structure can be experimentally determined using methods such as X-ray crystallography. That requires the purification and crystallization of the protein. Electron microscopy does not require a crystal, but there are few high resolution structures in PDB.

5 Structural changes upon ligand binding

(Brylinski and Skolnick, 2007) investigated the magnitude of structural changes in a protein resulting from ligand binding. They gathered the structures from PDB with resolution better or equal to 2.5 Angstroms, classified protein chains into apo and holo forms of the protein molecule and paired the chains at 100% sequence identity of the contiguous fragment of observed residues, i.e. residues for which backbone coordinates were determined. To remove redundancy, they clustered these sequences using a cutoff of 35% between clusters. Resulting in a dataset of 521 pairs of comparable apo and holo structures. They used a program to identify individual protein domains based on the structure.
The secondary structure, as classified to 8 types by the Dictionary of protein secondary structure (DSSP) (Kabsch and Sander, 1983), on average stayed similar upon ligand binding (around 95% identity between apo and holo forms).
RMSD, a measurement of global structure similarity, of individual protein domains in multiple-domain chains was smaller than the RMSD of single-domain chains. Therefore, packed individual protein domains are less sensitive to the state of ligand binding. However, in proteins with large RMSDs (>1 Angstrom), the multiple-domain proteins are overrepresented compared to single-domain proteins. This as well as the observed stability of individual domains can be explained by the movement of the entire domains wrt. each other in those high-RMSD multiple-domain chains.

6 Software framework

In addition to the software work itself, we present the results of selected analyses of the current version of the data in the PDB in the following chapters.

We also verify the results against previous work by (Brylinski and Skolnick, 2007).

6.1 Tools

6.1.1 Protein Data Bank

Protein Data Bank (PDB) is a database of experimentally resolved protein or nucleic acid structures, i.e. their modeled 3D shape. The structures are deposited by structural biologists in the form of PDB entries, which constitute the database. Various experimental methods are employed to obtain a structure. The database is updated weekly with new validated entries.

The metadata such as experimental method, polypeptide sequences, as well as the actual atom coordinates are publicly available in text-based mmCIF format.

6.1.2 UniProt

The Universal Protein Resource (UniProt) (The UniProt Consortium, 2021) is a resource for protein sequence and annotation data (namely cross-references to genetic databases, citations, protein name and its function, organism taxonomy). It sources protein sequences from translated genetic sequence data as well as from PDB. It consists of three databases, UniProtKB (subdivided to Swiss-Prot and TrEMBL), UniParc, and UniRef, which contain protein sequence data at different levels of non-redundancy.

UniProtKB/TrEMBL has one record for each full-length sequence in one species. Thus protein fragments of different length or isoforms produced by alternative splicing have different entries. However, it eliminates redundancy of identical sequences, across and within different sources. This part of the database is updated automatically from its sources.

UniProtKB/SwissProt has one record per gene in one species. It integrates all protein products of one gene. The SwissProt entry is created or updated manually from new TrEMBL entries which are subsequently removed.

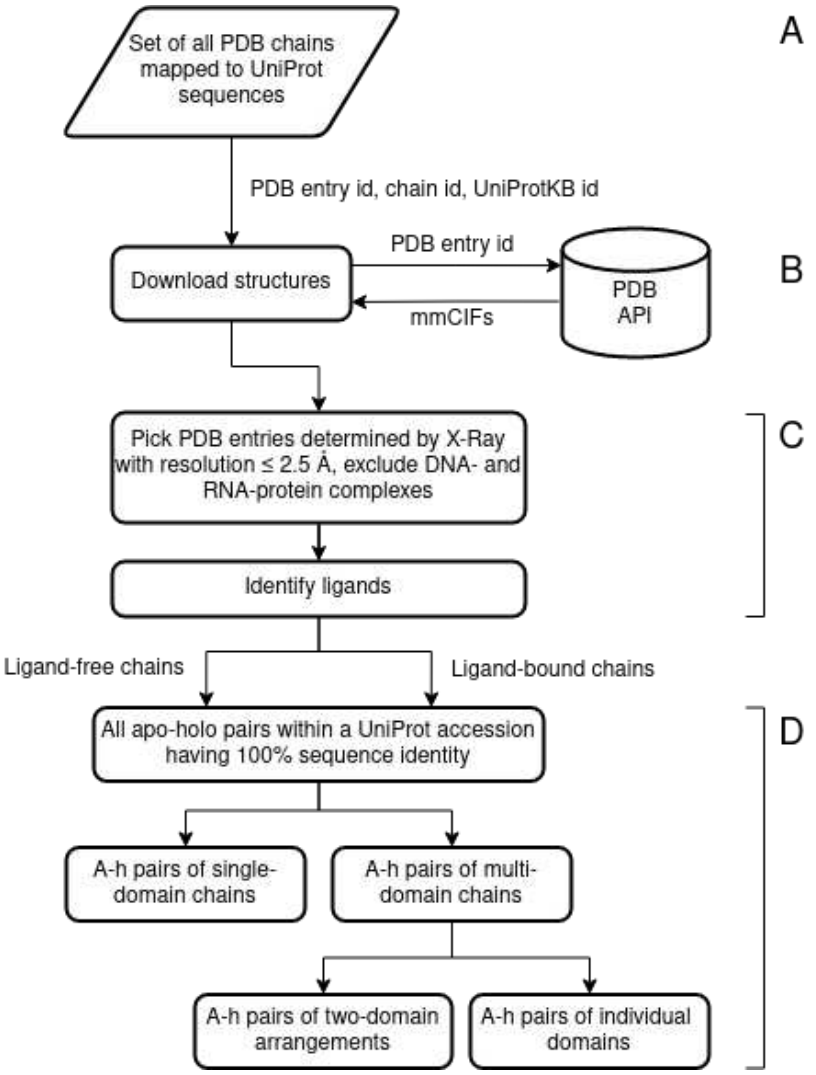
UniParc (UniProt Archive) is similar to UniProtKB/TrEMBL, in that it has one record for each full-length sequence, however, regardless of species. It also contains sequences that are excluded from UniProtKB, such as synthetic sequences or proteomes identified as highly redundant (e.g. thousands of bacterial strains).

UniRef (UniProt Reference Clusters) has one record for a sequence *and* possibly its shorter fragments, regardless of species. The member sequences of each entry (UniRef100 cluster) have ungapped local alignment of 100% sequence identity with the longest sequence in the cluster called “seed sequence” (Holm and Sander, 1998; Suzek et al., 2007). By further clustering the seed sequences at lower identity threshold, it provides clusters with minimum sequence identity of 90%, UniRef90, or 50%, UniRef50. (“About UniProt,” n.d.; “How redundant are the UniProt databases?,” n.d.)

6.2 Workflow

The largest entity we compare in ligand-free and ligand-bound forms is a protein chain. While we could also compare structures on the level of protein complexes (multimers), we leave it for future work. At the same time, we don't account for interactions with other polypeptide chains in the PDB structure.

First, we retrieve a file that lists all chains in the current PDB with its mappings to UniProt sequences; an output of SIFTS process (Dana et al., 2019) available in PDBe-KB. Three purposes are served – we obtain a list of (nearly) all PDB entries, each we subsequently download and process, and we can use the assignment to UniProtKB reference sequence in later steps, to reduce computation, as well as to subsample our output dataset in order to interpret the results.



We only consider chains in resolved structures not containing polynucleotide chains, with minimum length of 50 amino acid residues and with sufficient resolution same or better than 2.5 Ångstroms, as in (Brylinski and Skolnick, 2007).

Polypeptide chains meeting above criteria are then classified as ligand-free or ligand-bound. As ligands we mean either groups of heteroatoms in a residue with a single `auth_seq_id` (an mmCIF data item) with minimum of 6 non-hydrogen atoms (to exclude salts, water, etc.), or a polypeptide chain with length at most 15 residues. All ligands in a resolved structure are identified. A polypeptide-chain is classified as ligand-bound, if at least four of its residues are within 4.5 Å of a ligand (closest atom pairs residue—ligand are examined). Otherwise, the chain is classified as ligand-free.

Next, the chains are grouped by its UniProt primary accession, as assigned by SIFTS process. All possible pairs within a UniProt group are then paired at 100% sequence identity, meaning

that when aligned according to their longest common substring (LCS), there are no mismatches (leading or trailing after the LCS).

The longest common substring of two sequences in a pair then yields a residue-level mapping between the apo and the holo structure, which allows for direct comparison of apo and holo structures.

Finally, we compared the apo and holo structures with focus on large-scale domain movements upon ligand binding, using similar analyses to those in (Brylinski, Skolnick). For example, for an apo-holo pair, we compared the identity of the secondary structure and RMSD of the chains, measured the translation and angle of domain motions (how, upon ligand binding, they moved w.r.t. each other), and measured the change in interface area between each neighboring domain..

6.3 Implementation

The multi-step process or *pipeline*, described in the previous chapter, can be executed in several scripts, passing to each the output of the previous steps. One of the reasons to split the functionality into multiple scripts is to reasonably manage resources – some scripts fetch network resources from APIs (mmCIF files, domain definitions, secondary structure), do not require many computational resources and the API requests can be parallelized to a certain extend (the APIs rely on user restraint), while other scripts do CPU-intensive tasks, namely parsing mmCIF files and constructing BioPython's `Structure` object, or computing molecular surface, and can run in multiple instances on any number of computational nodes.

The first step is to download structure files from PDB. We use an API (rcsb.org/..) that allows us to download individual gzipped structure files with concurrent HTTP requests. Beforehand, we obtain a list of all PDB structures and their polypeptide chains using csv outputs of EBI's tool SIFTS (Dana et al., 2019), including the UniProt accession the chain is mapped to. This information is helpful in constructing groups of the same molecules from which apo- and holo-form pairs can be obtained.

The outputs of the pipeline scripts are in JSON. We chose it for its simplicity, widespread use and availability of parsers in many programming languages. Python's standard library, however, does not have a built-in way to output a generator (a stream) of objects as they are created, so all program outputs need to be in `list` in memory before outputting them to a file. This could be a problem in script `run_analyses.py`, which has a rather large output (the JSON file would take ~1.6 GB, the list before dumping to file could take more than ten gigabytes). We did not experience the problem, as we ran the script in multiple instances of small batches of the input data and pooled the results afterwards. Another solution would be to swap the serializing method for a csv writer, or adapting the JSON encoder to work with streams of data.
Diagram/posloupnost skriptů (názvů), začátek této sekce.

In script `make_pairs_lcs.py` we pair apo and holo of equivalent protein chains. We use the PDB chain to UniProtKB sequence mapping to reduce the number of possible apo-holo pairs, for which longest common substring (LCS) would be computed. Only pairs of chains mapped to the same UniProtKB sequence are considered. This made the computation tractable and resulted in LCS computation for ~3M pairs.

6.4 Future options

6.4.1 Protein complexes

The largest entity we compare in ligand-free and ligand-bound forms is a protein chain. We could also compare structures of protein complexes (multimers) as a whole. Unlike protein domains, which can be identified as spans of residues in a single sequence – protein chain – (so we can then find the same domain in the paired chain), there is no intrinsic ordering of monomers in a protein complex. To obtain residue–residue mapping between the residues of all the chains in both structures, the chains in apo and holo structures perhaps might need to be identified spatially, in case they are of the same sequence, for example by superimposing the two structure forms onto each other.
The information whether the studied molecule forms a complex is available in the PDB entry (section biological assembly).

7 Results

7.1 Datasets

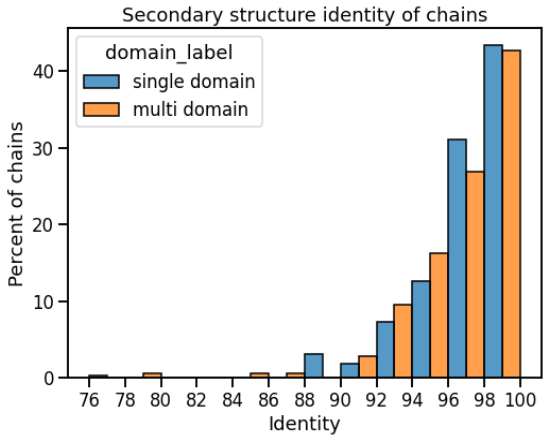
First, we compare the results of our pipeline with the results of previous work, to assert consistency. The non-redundant dataset (Brylinski and Skolnick, 2007), compiled from PDB as of October 2006, contains 521 chain pairs.

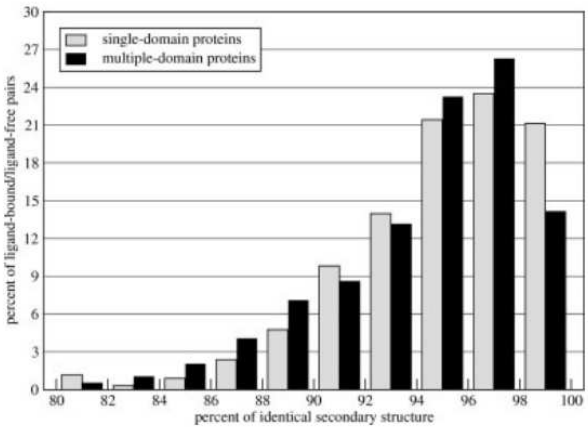
Next, the pipeline is run on the current dataset of 597,237 chains, of 170,910 unique structures from PDB. For interpreting the results, we selected one pair per unique Uniprot primary accession, removing the redundancy, resulting in 4674 pairs.

7.2 Comparison with previous work

(Brylinski and Skolnick, 2007) provide the non-redundant dataset of 521 chain pairs on their website (“The global structures of apo and holo proteins,” n.d.). They identified 22 thousand protein chain structures in PDB (October 2006), meeting the same criteria as described in above Methods chapter, 60% of those were classified as ligand-free structures, the remainder as ligand-bound. Fragments of protein chain sequences, having at least backbone coordinates, were then paired at 100% sequence identity resulting in 25,344 apo-holo pairs (all possible

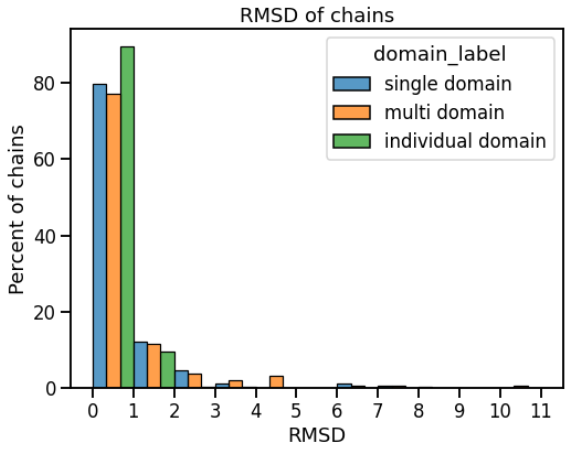
combinations). The sequences of the pairs were subsequently clustered at 35% sequence identity cutoff, yielding the 521 representative apo-holo chain (fragment) pairs. For a more detailed description we refer the reader to the paper. (For example, ligand-bound classification method differs from ours.)
We compared the classification to apo and holo forms, the pairing of the structures, and the results of the analyses of structural change. Input to our pipeline was, the dataset's, in total, 1042 structures. 1032 passed without errors, such as presence of a polynucleotide chain in the structure, or microheterogeneity in sequence. Of those, 95% were classified accurately, with 6 falsely as holo and 45 falsely as apo, wrt. results of (Brylinski and Skolnick, 2007). Pairing of those structures resulted in 464 same pairs as in their work, yielding recall of 89%.

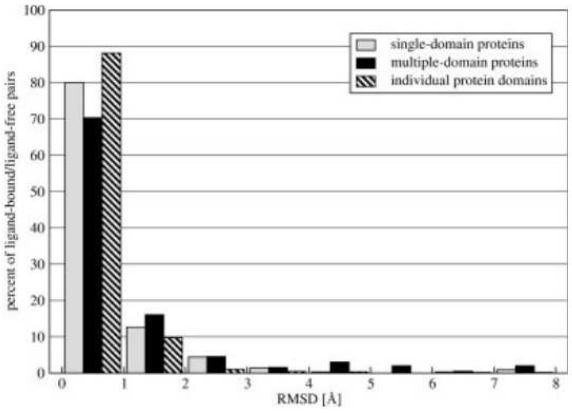




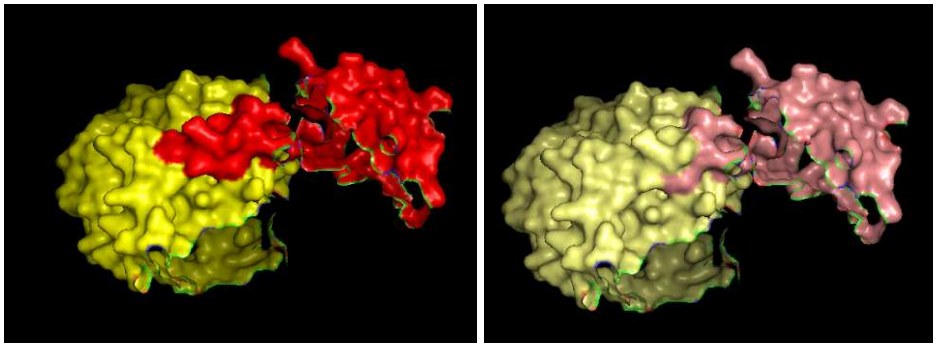
Ours Theirs

The difference in the distribution of secondary structure identity is caused by the fact that (Brylinski and Skolnick, 2007) used 8 types of secondary structure, while we had used classification to 3 types (alpha helices, beta sheet and none).



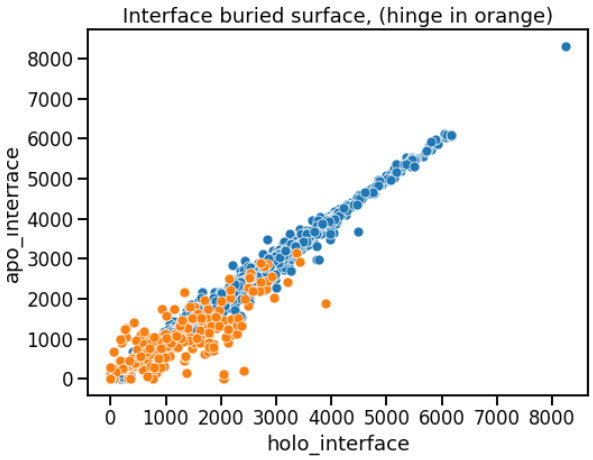


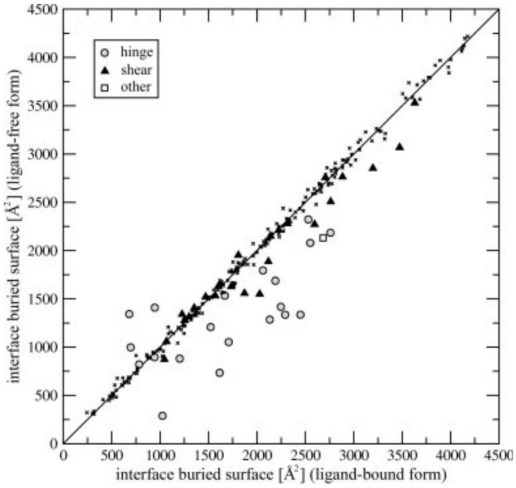
Despite different methods in the apo/holo classification and pairing the chains, the end result of pairing as well as the results of the analyses are similar and therefore it enables us to compare the results on the current PDB.



Two-domain arrangement. On the left domain boundaries visualized on the surface of the protein in our results, on the right domain boundaries of (Brylinski and Skolnick, 2007). This difference contributes to a larger measured domain interface area in our method.

7.3 Results on recent dataset





We observe there is a clear relationship between the magnitude of change in domain interface area and the absolute (apo) domain interface. In the extremes (largest changes for each imagined bin of absolute interface area), it appears almost linear.

8 Conclusion

We developed a software framework for analyzing apo and holo forms of protein structures. We successfully verified it against previous work and ran it on the current version of PDB. We obtained similar results, some phenomena like the dependence of the magnitude of domain interface area change on the absolute (apo) domain interface are visible more clearly, as there are more protein structures.

9 References

About UniProt [WWW Document], n.d. URL https://www.uniprot.org/help/about (accessed 1.6.22).

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. https://doi.org/10.1093/nar/28.1.235

Brylinski, M., Skolnick, J., 2007. What is the relationship between the global structures of apo and holo proteins? Proteins Struct. Funct. Bioinforma. 70, 363–377. https://doi.org/10.1002/prot.21510

Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., Velankar, S., 2019. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. 47, D482–D489. https://doi.org/10.1093/nar/gky1114

Holm, L., Sander, C., 1998. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14, 423–429. https://doi.org/10.1093/bioinformatics/14.5.423

How redundant are the UniProt databases? [WWW Document], n.d. URL https://www.uniprot.org/help/redundancy (accessed 1.6.22).

Huang, Z.-L., Urade, Y., Hayaishi, O., 2011. The role of adenosine in the regulation of sleep. Curr. Top. Med. Chem. 11, 1047–1057. https://doi.org/10.2174/156802611795347654

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637. https://doi.org/10.1002/bip.360221211

Morton, W.M., Ayscough, K.R., McLaughlin, P.J., 2000. Latrunculin alters the actin-monomer subunit interface to prevent polymerization. Nat. Cell Biol. 2, 376–378. https://doi.org/10.1038/35014075

Snyder, S.H., Katims, J.J., Annau, Z., Bruns, R.F., Daly, J.W., 1981. Adenosine receptors and behavioral actions of methylxanthines. Proc. Natl. Acad. Sci. 78, 3260–3264. https://doi.org/10.1073/pnas.78.5.3260

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

The global structures of apo and holo proteins [WWW Document], n.d. URL https://sites.gatech.edu/cssb/ligandbinding/ (accessed 3.6.22).

The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100