Charles University

Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Adam Král

Framework for retrieval and analysis of protein 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 also verify results against previous work on an equivalent dataset, and obtain results for the current version of PDB.

Keywords: protein; structural bioinformatics; PDB

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 na datasetu z výchozí výzkumu a ověřujeme je. Získáváme výsledky na aktuální verzi PDB.

Klíčová slova: protein; strukturní bioinformatika; PDB

Table of contents

1 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, and are thus 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 a protein of interest could bind a ligand, we are presented with an apo structure. To evaluate an algorithm for binding site prediction, we need a dataset of apo and holo forms of proteins (apo-holo pairs), for which their structure is resolved. Then we can check the prediction of the binding site in the apo structure and experimental evidence in the corresponding holo form. Or, for the same task, we might want to collect 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 on just holo structures – on prediction effectively treating the apo structure as a holo structure minus the ligand – would not detect the binding site; there would be no cavity for the ligand visible in the screened apo structure. 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). Beforehand we will introduce important terms and processes regarding ligand binding and protein structure (Chapter 2), which will build an understanding necessary for the following chapters. We will also describe specifically the protein structural changes induced by ligand binding (Chapter 3).

2 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 drugs but also (natural) substrates to enzymes, signaling molecules, etc.

Binding a ligand may induce a conformational change in the protein (i.e. change its 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) show the ligand is deeper in the actin structure and from there it deforms the subunit interface, which hinders the association of the subunits. This is an example of allostery; where the activity of a protein is affected by a ligand binding to a distinct site from the site responsible for the protein function.

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).

2.1 Protein structure

Proteins consist, in particular, of one or more linear chains of amino-acid residues. The chain is a polymer where the residues are joined with a peptide bond, hence also termed as polypeptide. Proteins of multiple chains are also referred to as protein complexes.

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 (see Figure 0; which is itself part of the backbone). The chain is somewhat flexible, as bonds around the alpha carbon generally allow more than one *torsion* (also *dihedral*) *angle*. Preferred bond angles of residues from a non-redundant dataset of 500 structures (Lovell et al., 2003) are shown in Figure 1

Protein structure can be described on multiple levels – so-called primary, secondary, tertiary and quaternary structure; in the order from more local to global phenomena.

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

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

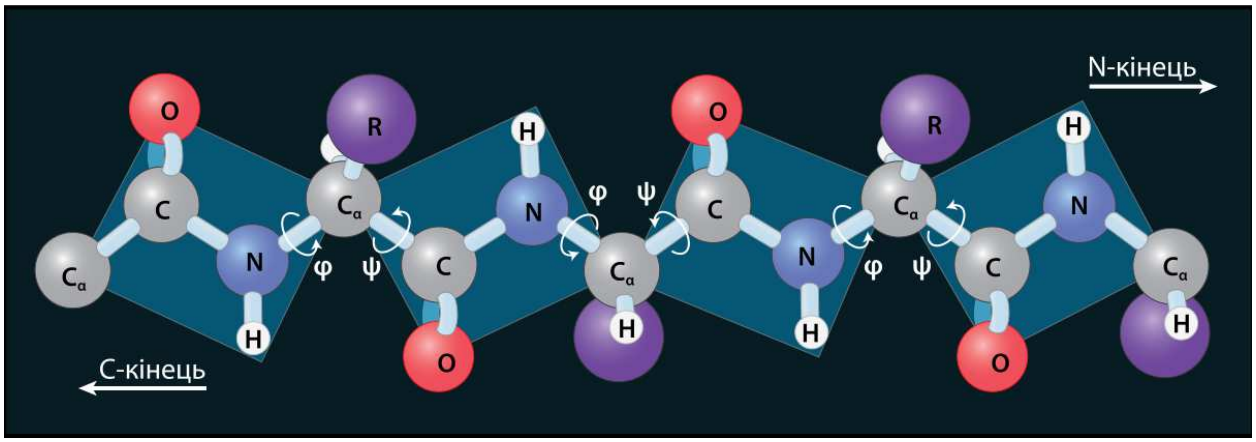


Figure 0. Peptide, a chain of amino acid residues. Purple spheres with label R stand for amino acid side chains, which is a group of atoms specific for each amino acid. The peptide bond is planar (C=O and C-N bonds lie on the same plane; the C-N bond has a rigid, partial double bond character, emphasized by the plane in the figure) and the bonds that are rather free to rotate around in the protein backbone are the N-Cα bond (Φ torsion angle in the figure) and the Cα-C-carbonyl bond (ψ torsion angle) in each residue. Figure by (Zlir'a, 2013)

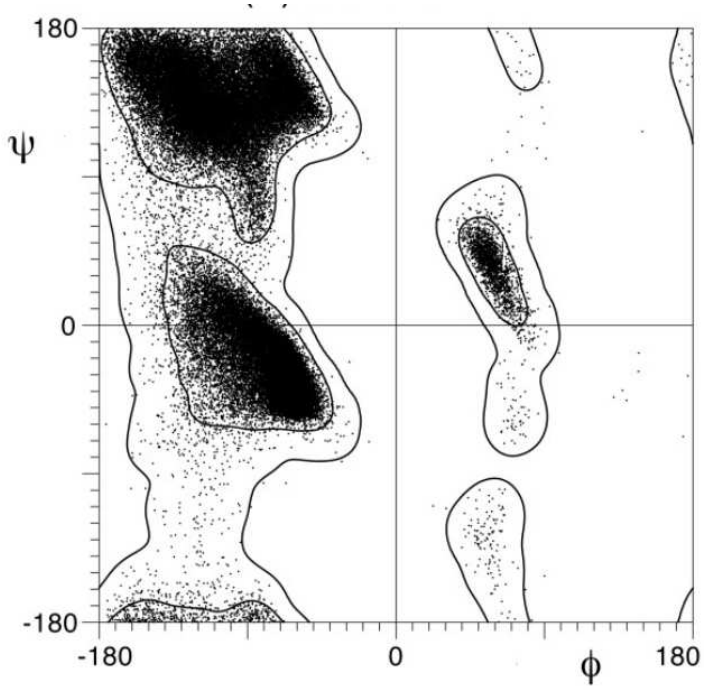


Figure 1. Torsion angles Φ and ψ around the alpha carbon of PDB residues (excluding glycine, proline and the residues preceding proline – those have very different plots). Figure by (Lovell et al., 2003). Some angles are unfavorable due to steric clashes (Lovell et al., 2003), ie. van der Waals repulsion of too close atoms.

Secondary structure is the local structural pattern of the residues, more precisely their backbone. There were observed different patterns (types) of secondary structure in proteins when their structures became resolved. A protein structure can contain multiple types of secondary structure, as well as multiple distinct segments of the same type. A popular formalization of the secondary structure, Dictionary of Protein Secondary Structure, and the method for automatically classifying it known as DSSP (Define Secondary Structure of Proteins) was developed by Kabsch and Sander in 1983. It assigns one of 8 classes to each residue in the protein chain. Instead of defining the structure types using the dihedral angles, which do completely describe the local backbone orientation, it uses the existence of hydrogen bonds in the backbone for the definitions (between carboxyl oxygen and amino hydrogen). The existence of hydrogen bonds is decided if the magnitude of the electrostatic energy is larger than a given threshold and the electrostatic energy is calculated using the coordinates of four atoms (carbon, oxygen, nitrogen, hydrogen – which is added in as it is typically not present in X-ray structures), taking into account the directionality of hydrogen bonds. As they argue, this leads to a simpler definition than setting bounds on the dihedral angles for the residues for each structural type.

For example there is a 4-turn at i-th residue, if there is a hydrogen bond between i-th and (i+4)-th residue. And there is an alpha helix in the span of (i, i+3) residues, if (i-1)-th and i-th

residues are 4-turns. Finally, the helix classification ('H' for alpha helix) takes precedence over the turns (3,4,5 turns are assigned class 'T') and one class is reported for each residue.

The reported classes are 3,4,5-turn helices (G, H, I), beta strands (E, or B for a *single*-residue bridge), T for the turns, S for a bend (which is by exception computed using angles between two vectors, a vector from alpha carbon i-2 to i, and the other vector from alpha carbon i to i+2; it is a bend if the angle is larger than 70°), and C for a random coil (none of the previous classes assigned).

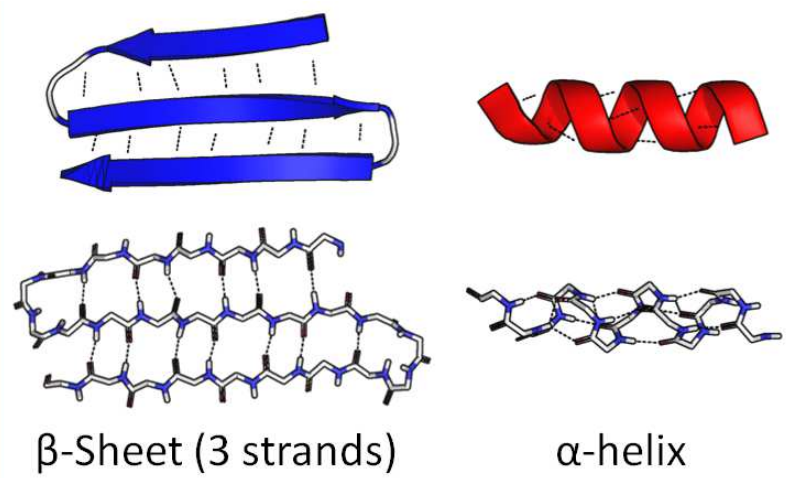


Figure 2. Two most common types of secondary structure - alpha helices and beta sheets in their cartoon representations (top) and with atom sticks (bottom). Dashed lines show stabilizing hydrogen bonds. (Shafee, 2017)

The overall chain's three-dimensional fold is called the *tertiary structure*. (And *quaternary structure* involves the association of multiple protein chains, subunits, to form a protein complex.)

2.2 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 they may undergo movements relative to each other.

2.3 Determining protein structure

Protein structure can be experimentally determined using methods such as X-ray crystallography. That requires crystallization of the protein. That is done through trial-and-error, by setting up an array of different chemical, biochemical and physical conditions the protein is exposed to, until it precipitates and forms a crystal appropriate for the X-ray diffraction analysis.

Moreover membrane proteins, with both hydrophilic and hydrophobic surfaces (in vivo, cytosolic/extracellular parts vs. transmembrane parts), are difficult to crystallize. (McPherson and Gavira, 2013)

Electron microscopy (EM) does not require a crystal, but there are few high resolution structures in PDB. The output of both methods is a reconstructed 3D electron density map, in which the atomic model is fitted. Whereas single-particle EM has a potential to retrieve multiple conformations of the protein (Cheng et al., 2015), because the data is collected for individual molecules, the density map in X-ray crystallography is already averaged across the conformations of the molecules in the crystal and most methods return a single atomic model (Keedy et al., 2015).

Nuclear magnetic resonance spectroscopy (NMR) gives rather local constraints on the structure; the observed electromagnetic (radio frequency) spectra contain signal peaks emitted from the molecules' nuclei at frequencies that depend on the atom and its chemical surroundings.
So-called spatial restraints are obtained from the measured data, after the data being mapped to each residue; position in the protein sequence. The overall structure can be reconstructed by energy minimization of the structure using the restraints (and modeled natural attractive and repulsive forces). The minimization is performed from multiple random initializations of the structure, returning an *ensemble* of structures. Molecular dynamics with simulated annealing is often used for the optimization, where the energy of the protein (atoms' kinetic energy) is increased after reaching local minima in the optimization process. For there are multiple local minima of the conformational landscape of proteins; with energy barriers to switch in between them. (Marion, 2013) Arising in the computation, but also experimentally observed (Ulrich Nienhaus et al., 1997). Note that in general, the energy barrier does not mean the values of the local minima have to be different, both conformations can be equally stable.
More recently, low populated conformational states and protein dynamics (also in cells in vivo) can be studied using solution NMR (Hu et al., 2021), but not using EM or X-ray crystallography where the sample is solid.

However, it is difficult to solve larger structures using NMR (Marion, 2013), the median deposited model molecular weight (excluding waters) in PDB is 10 kDa, while for X-ray structures it is 52 kDa, 7 % structures are solved using NMR, 87 % using X-ray diffraction (as of August 2022).

2.4 Protein dynamics and ligand binding

Proteins in a solution are not static despite it may seem like that from the list of coordinates of atoms in PDB.

The random zig-zag motion of a relatively large particle in water (pollen, observed by botanist Robert Brown, Brownian motion), was suggested as possible evidence of thermal motion of molecules by (Einstein, 1905) and subsequently verified.

The energy levels of particles (in a gas, if the particle interactions are negligible) follow a Boltzmann distribution. It is the most probable combinatorial division (of all applicable division) of a set of N particles into some discrete energy levels (possibly infinitely many) E_i, such that the sum of energies of the particles is a chosen constant, the total energy of the system; in the limit as the number of particles N goes to infinity. As a result, the probability of a particle at energy level E_i is proportional to $e^{-E_i\beta}$, where β is uniquely determined by the values of the energy levels and the total energy of the system. In thermodynamics, β is inversely proportional to the temperature of the gas. (Atkins and De Paula, 2006; also contains the derivation of Boltzmann distribution using Lagrange multipliers. The same formula, Boltzmann distribution, can also be derived for continuous states.)

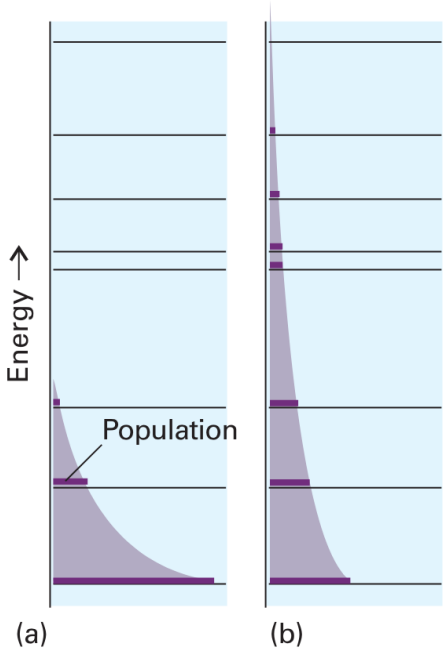


Figure 3. Boltzmann distribution for some energy levels, depicted as horizontal lines. The exponential function $e^{-E\beta}$ (in energy of the state, on the vertical axis) is shown continuous, although here the particles can populate only the discrete levels. Boltzmann distribution (b) has lower β (higher temperature) than (a). The discrete distribution (b) has a greater entropy than (a) (information-theoretic, as well as thermodynamic entropy of the corresponding system). Figure from (Atkins and De Paula, 2006).

It is not evident what the distribution of molecular energies would look like in the cell, liquid, or other more complex system consisting of multiple different compounds. Where intermolecular interactions introduce also the potential energies of the molecules and permit the exchange of variable amounts of energy between the molecules. Besides the potential energy, a component of the molecules' energy is the kinetic energy, manifesting in molecular translation, and particularly in liquids, vibration and rotation (where molecular vibrations would be the periodic motion of atoms relative to each other; the bond lengths and angles may not be static; Atkins and De Paula, 2006). However, it is easy to imagine that just like the pollen particle was moved,

when the contributions of the interaction of water molecules were uneven, the protein molecule, which is an order of magnitude smaller, will be even more subjected to those random fluctuations of the surroundings (including the presence of ions or other molecules). The relatively free-to-move-around bonds in the carbons of side chains and those next to the alpha carbon of the backbone, which determine the protein fold, can change their torsion angles, and the regions of the structure can even escape their stabilizing interactions with other atoms, depending on the energy barrier to break the interaction. Therefore the protein structure can have multiple shapes, forms, at a given time and could occupy multiple, even radically different, conformational minima simultaneously in the solution.

Over time, various models of ligand protein interaction were formulated (Figure 4). In the lock-and-key model, the ligand shape should be complementary to the protein binding site. Induced-fit model was then developed for structures, which could not be explained by the lock-and-key model and hypothesized that the final protein shape is also determined by the ligand. Conformational selection model assumes a pre-existing equilibrium of multiple conformations, as described above, where the ligand will bind only to a specific one. (Keskin, 2007; Nusrat and Khan, 2018)

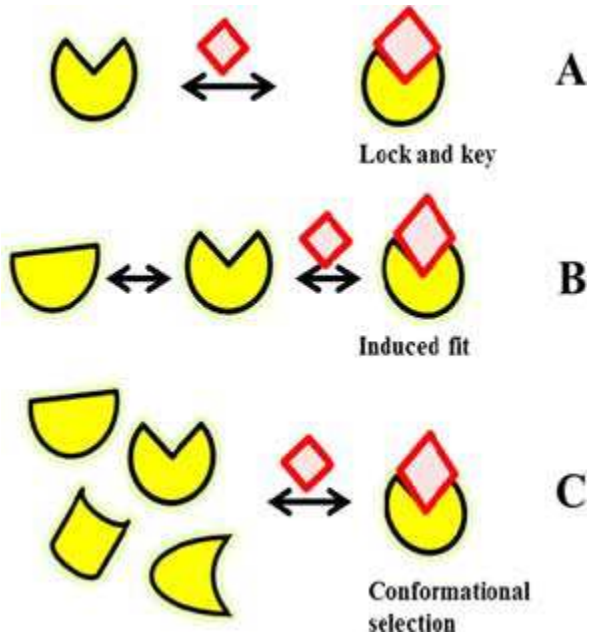


Figure 4. Models of molecular recognition of protein ligands. Lock-and-key model assumes no conformational changes upon ligand binding. The ligand (red) and the target (yellow) have complementary structures. Induced fit model, the conformation of target changes due to the ligand interaction. Conformational selection model, the native state is actually an ensemble of conformations, the deformations may occur even before the binding of ligands. Figure and caption from (Nusrat and Khan, 2018)

Electrostatic interactions between ligand and protein, hydrophobic interactions, van der Waals interactions (attractive and repulsive), pi-interactions are the most significant forces in ligand

binding (Ahmad et al., 2013; Atkins and De Paula, 2006, p. 638; Nusrat and Khan, 2018), their short description should follow.

However, just the existence of attractive forces cannot explain why the ligand would bind, and not be attracted and eventually repelled away again (by the van der Waals repulsion). It is the interactions with other molecules that exchange the energies in between the molecules, and both protein and ligand would be in solution in contact with the surrounding molecules. It is very improbable that during the approach and binding of the attracted ligand, none of the extra energy, converted from the potential energy due to the attraction, would be dissipated to the hundreds surrounding molecules and would remain only in the two - ligand and protein. A similar argument could be made for why the protein would be most likely to be found in its conformational local minima. Macroscopically this can be explained by (one of the formulations of) the second law of thermodynamics: “The entropy of an isolated system increases in the course of a spontaneous change” – (Atkins and De Paula, 2006). Therefore, after adding the ligand, the amount of bound ligand will spontaneously increase until reaching equilibrium, increasing the entropy.

Or if the system was not isolated, but at constant temperature and pressure (conditions for biochemical reactions), the law for isolated systems will still hold for the combination of the *system* and its *surroundings* (together forming an isolated system). Then the process will be spontaneous, as long as the change in Gibbs energy of the *system* would be negative. Where the change in Gibbs energy, at that conditions of constant temperature and pressure, balances the heat transferred to the surroundings (this increases the entropy of the surroundings) and the entropy change of the system.

$$\triangle G = Q - T\triangle S$$

Equation for the change of Gibbs energy at constant temperature T and pressure. For example, process can be spontaneous ($\triangle G < 0$) even if the entropy of the system decreases ($\triangle S < 0$), because it can be balanced by transferring heat to the surroundings ($Q < 0$); so that the entropy of the surroundings increases and the second law is not violated.

Isothermal titration calorimetry can determine the thermodynamic parameters of ligand binding such as the change in Gibbs energy, and the binding constant K_b. The binding constant

$$K_b = \frac{[ML]}{[M][L]}$$

where $[ML], [M], [L]$ are the concentrations of protein-ligand complex, free protein and free ligand respectively, and is a measure of how effectively the ligand binds the protein; its affinity.

The dependence of the change in Gibbs energy is given by $\triangle G = -RTlnK_b$, where R is the gas constant.

In drug design, the affinity of the ligand binding to the target is important for the drug to be effective. However, equally important is the affinity to all other off-target proteins as most drug candidates fail clinical trials due to side effects caused by the interactions with off-targets (Pinzi and Rastelli, 2019). By increasing the affinity to the target, lower doses can be administered, lowering the severity of the adverse effects.

As a side-note, *covalent* bonds between cysteine residues (disulfide bridges, “bridge” the residues from different locations in the primary structure) can fix a particular fold. Failure in formation can lead to incorrect folding, loss of the biological function, and protein aggregation. Disulfide bridges are mostly found in proteins in the extracellular environment. (Wiedemann et al., 2020). Perhaps because the conditions are more variable (such as pH) which can affect the strength of the non-covalent interactions.

3 Structural changes upon ligand binding

Here we focus on the structural changes that can be observed in two states of the structure – before and after the ligand binding, primarily from the crystallographic evidence. The process is in fact dynamic and may happen on a variety of timescales, depending on the complexity of the movement and the energy change, but that is a subject of protein dynamics studies which would use for example suitable NMR techniques (Greener et al., 2017).

It is hypothesized that for catalysis (enzymes) the conformational change should involve a low energy barrier (Gerstein et al., 1994) – that adds to the activation energy of the reaction, which influences the reaction speed – so that the reaction can be fast, and we should investigate primarily those movements. The author finds an exception to this rule, but it is an allosteric protein where the conformational change alters a subunit interface, quaternary structure, so there is not the pressure on fast conformational changes (Gerstein et al., 1994).

As mentioned in the previous chapter, the flexibility of the protein chain stems from the relative freedom in the torsion angles. However, we shall describe the movements on a larger scale and try to put them in perspective.

Structural changes include the movement of residue’s side-chains in the binding site, loops (regions without a regular secondary structure usually near the surface of the protein, connecting helices and sheets), and can induce also large movements on the scale of domains (Brylinski and Skolnick, 2007).

The change in the backbone torsion angles can be localized, or spread out in larger regions. For example loops or strands that are not restricted by hydrogen bonding in a sheet have a relative freedom in their torsion angles. Therefore for a large structural change to happen, only few localized residues need to change their torsion angles; the residues effectively forming a hinge. Whereas in helices the pattern of hydrogen bonding is more restrictive and to induce a

significant structural change many residues need to adjust the torsion angles. (Gerstein et al., 1994)

(Hayward, 1999) shows an example of an enzyme with a helix that is bent in the ligand-free conformation, and straight in the ligand-bound; it is believed to store elastic energy for the rapid closure of the ligand. As calculations show the energy of the hydrogen bonding in the bent helix larger than the straight. Note that this may conflict with the requirement of low energy barrier of the conformational change after ligand binding suggested earlier. However the barrier might be high, if the helix harnessed the energy of the product, to use it on the next substrate molecule to lower the activation energy; the high conformational energy barrier would be overcome with the released chemical energy. After the enzyme synthesis, the first reaction might need to happen without this potentiation though. (Which is not unthinkable, there are *some* high energy substrate molecules, see the chapter with Boltzmann distribution, although not enough for a fast sustained reaction. Or the enzyme may switch to the potentiated helix conformation without the substrate. Again acceptable one time, but not regularly during the reaction that should have a fast turnover of the substrate.)

Closely packed segments with many stabilizing contacts between residues arising from the tertiary structure preclude the backbone torsion angle changes. However a packed interface between two chain segments connected with a hinge *outside* the constraint packing can allow a motion either perpendicular to the interface – hinge motion (contacts are made and lost), or parallel to the interface – shear (contacts are usually preserved, only the side-chains may lean in the direction of the motion – undertaking a small change in *side-chain* torsion angles).

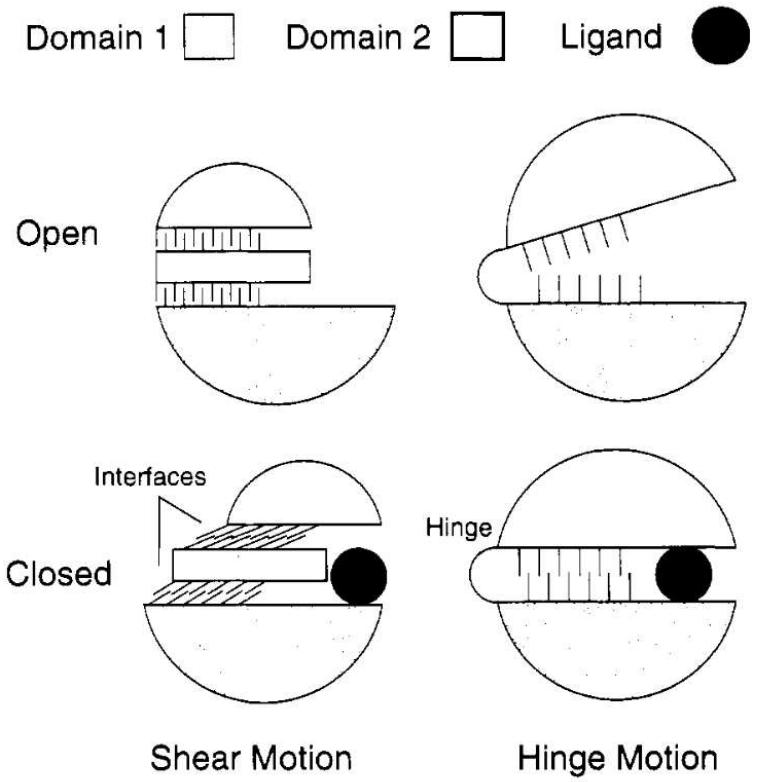


Figure 5. Hinge and shear motion. Example with ligand closure. Interface of packed side chains shown as interdigitating lines. In shear, the contacts of the interface are mostly preserved, and this results in a relatively smaller range of motion than the hinge motion on the right. However, multiple shear interfaces can add to the motion. Closure of the ligand may help the mechanism of catalysis of an enzyme, or can allosterically propagate the signal of binding a ligand. Figure by (Gerstein et al., 1994).

Domains are an example of such closely packed segments. Figure 5. shows the motion of two domains. The distinction of hinge and shear movements is for illustration, in reality the movements may have elements of both. The magnitude of shear motion on one interface is relatively constrained, but multiple shear interfaces can produce in total a large movement. (Gerstein et al., 1994) gives an example of citrate synthase, which exhibits multiple local shear movements on helix-helix interfaces of its small domain. 9 helices move relative to the neighboring by up to 1.8 angstrom and rotate by up to 13°. The helices themselves are more or less rigid. The composition of the movements results in the distant helix translating by 10 angstroms and rotating by 28°, compared to its ligand-bound position.

In some cases, changes may happen in the secondary structure. In pheromone-binding protein, a C-terminal helix occupies the binding site. During the binding of the ligand, bombykol, the entire helix unwinds, forming a loop (Brylinski and Skolnick, 2007).

An extreme example in the change in secondary and tertiary structure are some intrinsically disordered proteins (without a rigid fold, lacking a stable tertiary or even secondary structure) can become structured upon ligand binding (Ahmad et al., 2013).

3.1 A statistical study

(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 automatically 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 relative to each other in those high-RMSD multiple-domain chains.

4 Software framework

We built up on the statistical study (Brylinski and Skolnick, 2007) and implemented an open source framework that can automatically process the current state of PDB.

In the next chapters we outline the workflow and verify the results against previous work by (Brylinski and Skolnick, 2007).

4.1 Tools

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.

As of August 2022 PDB contains 194011 of structures.

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 a record for a sequence, and possibly other records for 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.)

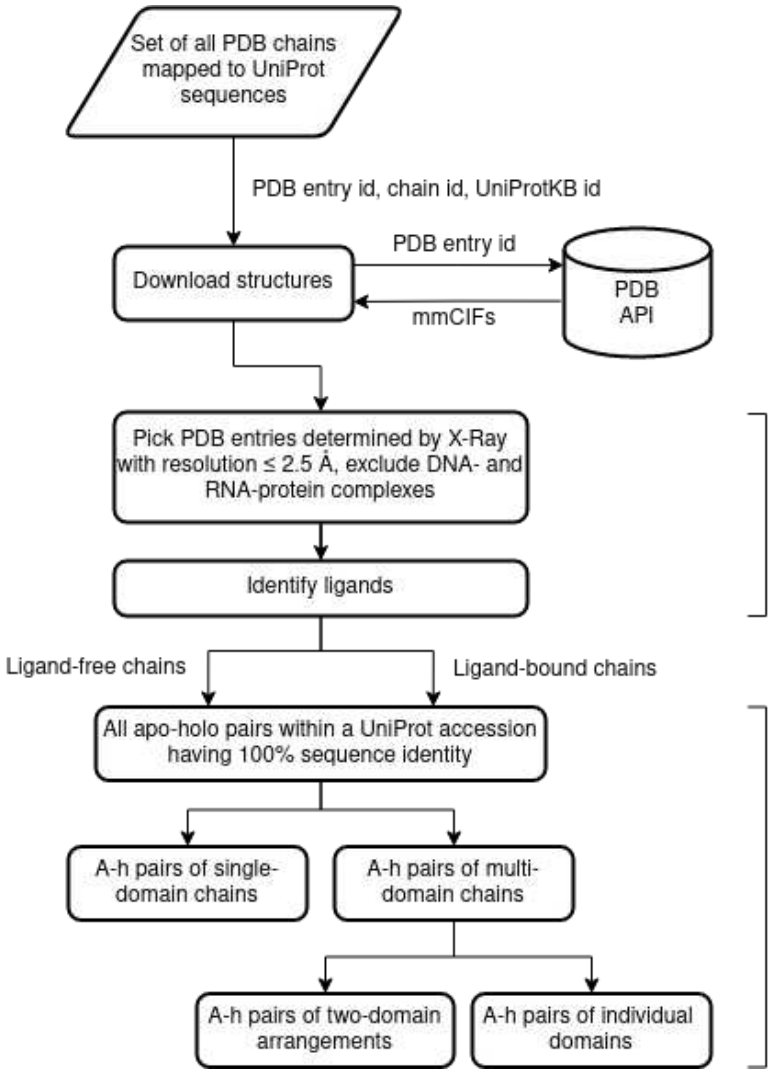
4.2 Workflow

Subject of study – chain

The largest entity we compare in ligand-free and ligand-bound forms is a protein chain, i.e. single polypeptide molecule. 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. (For the time being, we suppose this would not skew the results considerably and is done the same way in previous work, which did not comment on it. A subset of the output data can be selected which would not be subject to this problem.)

We also compare structures of individual domains in the chains.



Obtaining chains

First, we retrieve a file that lists all chains in current PDB with its mappings to UniProt sequences; an output of SIFTS process (Dana et al., 2019) available in PDBe-KB (PDBe-KB consortium, 2022). Multiple 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.

Eligible chains

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).

Determining ligand-binding state

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.

Pairing apo and holo chains with identical sequences

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.

Comparing structure of two chains

Finally, we compared the apo and holo structures with a 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 relative to each other), and measured the change in interface area between each neighboring domain.

Residue-level mapping

We need to be able to identify the same residue in both apo and holo forms. When computing the secondary structure similarity across the pair of chains, we would determine for a residue, if it is a part helix, say as in the other chain, or it is now classified as a coil and thus the secondary structure is different at this position. Or, for the computation of RMSD, we need to know the one-to-one map between points of the apo and holo structure (in our case the c_alpha coordinates).

We use the residue-level mapping from longest common substring computation and for all the subsequent structural analyses we choose *a subset of the longest common substring* - those residues, that were observed both in apo and holo structure (i.e. they had an atom coordinate determined for them).

Secondary structure similarity

We use PDBe-KB API to obtain the secondary structure. It provides the segments of residues forming helices and strands in the requested PDB entry. Other residues are classified as a coil, therefore yielding a total of three classes of secondary structure. The computed similarity (identity) is the percentage of residues having the same class in both the apo form and the holo form of the chain.

RMSD

We compute the RMSD across the apo and holo form using the Kabsch algorithm (Kabsch, 1976) (to align the two centered structures) available in the python package *rmsd* (Kromann, 2022). We use only the C_alpha coordinates in the computation (if not available, then any residue's atom coordinate).

Domain definitions

The domain definitions are obtained from a PDBe-KB API exposing CATH domain boundaries. Because the domain definitions don't always perfectly correspond in the two structures, we only use the boundaries of the apo chain domains and remap them to the holo chain (using the residue-level mapping).
We can then compare the structural characteristics not only across entire chains, but also across the individual domains (the chain's subset), or across *two-domain arrangements* which will be explained in the following section.

Interdomain surface area

The interdomain surface area (also interface buried surface area) is the area that is not solvent accessible between two domains in a structure – on their *interface* – but would be accessible, if the domains were solitary. And exactly in that manner it is calculated. First, the solvent accessible surface areas (SASA) are calculated for both domains, each placed in void, solitary, these numbers are added, then the SASA of the two domains (a subset of the original structure containing only the two domains) is subtracted. Yielding the buried interface area.
For calculation of SASA we use FreeSASA (Mitternacht, 2016), the input being the (sub) structures specified above.

Two-domain arrangement

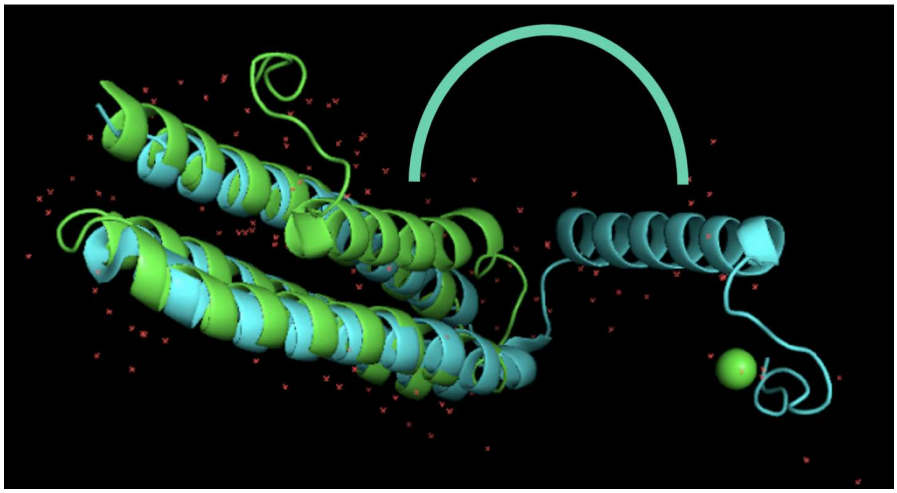
To measure a domain's movement after ligand binding we need a common reference frame in the two – apo and holo – structures. The natural choice is another domain. Then we get an easily interpretable result – compared with the reference frame being the entire structure, or all other domains. The interpretation follows – we forget about everything in the structures but the two domains in the apo structure and the same two domains in the holo structure (see Domain definitions). Then we observe how this domain pair differs in the two structures, and can use one of the domains as a common reference frame (explained in the next chapter) to calculate the movements, i.e. the change in relative orientation of the two domains.

We therefore get a movement measurement for one domain and the second domain being the reference frame. This unit of two domains is called a two-domain arrangement, using the terminology in (Brylinski and Skolnick, 2007). We can get the measurement for all possible pairs of domains in the molecule.

Describing the domain movement

To assess how the domain rotated with respect to another domain we can employ superposition with the Kabsch algorithm. We can align the two-domain arrangement in both structures by the *first* domain. Then we can obtain the *translation vector* of the *second* domain after ligand binding by subtracting the centroids of the domain in the apo and holo structure. Finally we can take these modified structures and obtain the *rotation matrix* for the second domain, by running the Kabsch algorithm again, this time superimposing the second domain. From the matrix, we can derive the *rotation angle*. If we consider only the absolute value of the rotation angle and the (positive) norm of the translation vector, the result does not depend on which domain from the two we choose as the *first* and the *second* (symmetry).

.



Relative movements of domains.
180° rotation of the helix domain. In blue the apo structure (1k04A), in green the holo (1ow6C). Structures are superimposed using the other domain.

4.3 Implementation

The repository (available at https://github.com/adam-kral/apo-holo-protein-structure-stats) contains a standalone pip-installable package of the software framework, as well as notebooks analyzing the results. The software framework package is divided into more packages but notably contains a package *pipeline* which consists of six scripts implementing the multi-step process or *pipeline* described in the previous chapter.

The scripts are structured as follows. The input to a script are the outputs of the preceding steps (scripts). 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 may be parallelized *only to a certain extent* (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. Another reason for multiple scripts is that a user can run any

pipeline step, perhaps with different settings, with the data they already have, or modify the outputs before executing the next step.

The scripts are installed for convenience (in a standard way) with the package into the python environment, so they can be executed using their aliases from anywhere in the system (as long as the python environment is in the *PATH* environment variable).

The list of pipeline scripts and their short description follows. The listed names are the aliases, but they mimic the names of the files in *pipeline* package that contain the definitions (code).

ah-chains-uniprot
Lists *all* PDB chains with their respective UniProt accession (for the polypeptide sequence). In case a list of PDB chains is already supplied by the user, it annotates it with the corresponding UniProt accessions.
ah-download-structures
Downloads the PDB structures (specified as a JSON list of chains) in a specified number of download threads.
ah-filter-structures
For each specified chain it opens the downloaded structure mmCIF file containing the chain and checks if the structure meets predefined criteria (in settings), otherwise it does not include it in its output. It also exploits the fact that the structure is already parsed in memory and annotates the chain with information obtained from the parsed structure file, such as classification of the chain to apo/holo form, and the chain sequence.
ah-make-pairs
It groups the chains by its UniProt accession and for all possible apo-holo chain pairings within the group it computes the longest common substring of the chain sequences. It also reports the number of mismatches preceding and trailing the LCS. In case they are non-zero, the chains are not deemed equivalent and their structure will not be compared in the next step.
ah-run-1struct-analyses
The *ah-make-pairs* script may greatly reduce the number of structures passing to the structural comparisons, therefore this script runs only analyses for the structures that occur in a valid apo-holo pair. Currently, it fetches secondary structure assignments and domain definitions from PDBe-KB API for each chain.
ah-run-analyses
Runs the structural comparisons of the apo-holo chain pairs. Loads the structure files and accesses the domain definitions and secondary structure assignments. Outputs the results of the analyses as well as the observed residues (positions) used in the comparison.

Output format

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. (Except for the script *ah-run-1struct-analyses* which caches the API responses using *shelve* standard library module (persistent dictionary with values as *pickle*d objects) to a dbm file.)

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 *ah-run-analyses*, which has a rather large output (the JSON file would take ~1.6 GB, that translates into the list before dumping to file taking up more than ten gigabytes of RAM). 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.

Notes

ah-chains-uniprot
The list of all PDB structures and their polypeptide chains with their UniProt accession is available in *csv* at
ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/flatfiles/csv/pdb_chain_uniprot.csv.gz.

ah-download-structures
We use an API (https://files.rcsb.org/download/{pdb_code}.cif.gz) that allows us to download individual gzipped structure files with concurrent HTTP requests (unlike the ftp access, available in BioPython Bio.PDB.PDBList).

In script *ah-make-pairs* 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. The non-redundancy of UniProtKB entries, as mentioned earlier, ensures all …

4.4 Future options

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).

5 Results

5.1 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 (see the results in chapter 5.2).

Next, the pipeline is run on the current dataset (as of 04/29/2022) of 595,878 chains, of 170,634 unique structures from PDB. For interpreting the results, we selected one pair per unique Uniprot primary accession, removing the redundancy, resulting in 4674 chain pairs (see the results in chapter 5.3).

5.2 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%.

Below, we show most of the analyses done in the paper (on the right) alongside our results (left). Please keep in mind that 11% of the pairs are missing in our results (see the *recall* above). Given the dataset was *non-redundant*, it could skew the results considerably (e.g. in histogram plots). Plots based on the *entire current* PDB (next chapter) look more similar to the plots from the paper in some cases (don't suffer from the skewing due to the important parts of the dataset missing.

Structural similarity

Here we show the secondary structure similarity (Figure 7) and RMSD across the apo and holo forms. The dataset of chain pairs is further divided into single-domain chains and multiple-domain chains. Also besides whole chains, structure of individual domains is compared, in case the chain is multiple-domain. This is the third dataset in the RMSD plot (Figure 8).

Different domain boundary assignment - we use CATH domain boundaries, which are obtained using a mixture of automatic methods and manual validation (“CATH,” n.d.; Greene et al., 2007) whereas (Brylinski and Skolnick, 2007) used an automatic method “Protein Domain Parser” (Alexandrov and Shindyalov, 2003) which, among other things, recursively splits chain into domains based on the number of residue contacts between the newly formed domains.

We did not compare the number of domains found in the chains (however it is available in the authors’ dataset), although we compared domain boundaries, which were for selected chains provided in the paper, on a case-by-case basis – if the results of the analyses differed considerably.

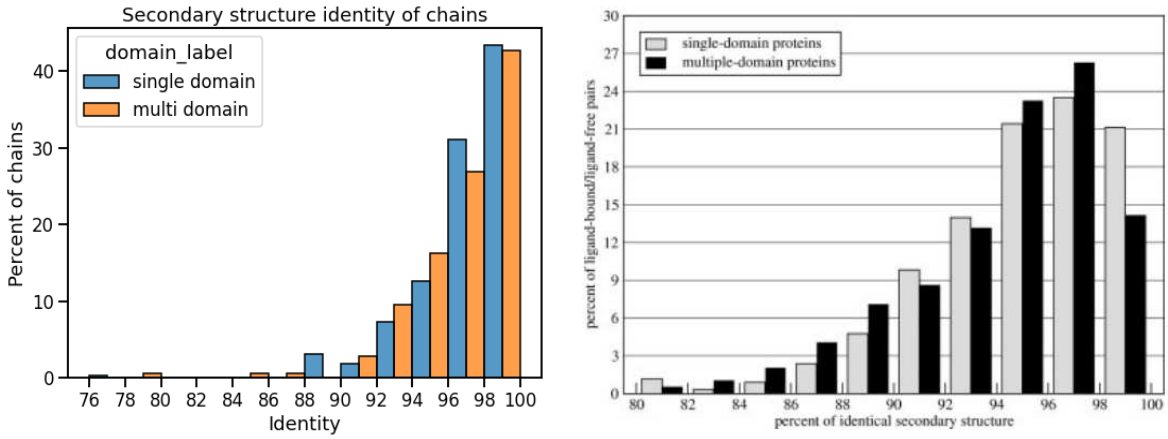
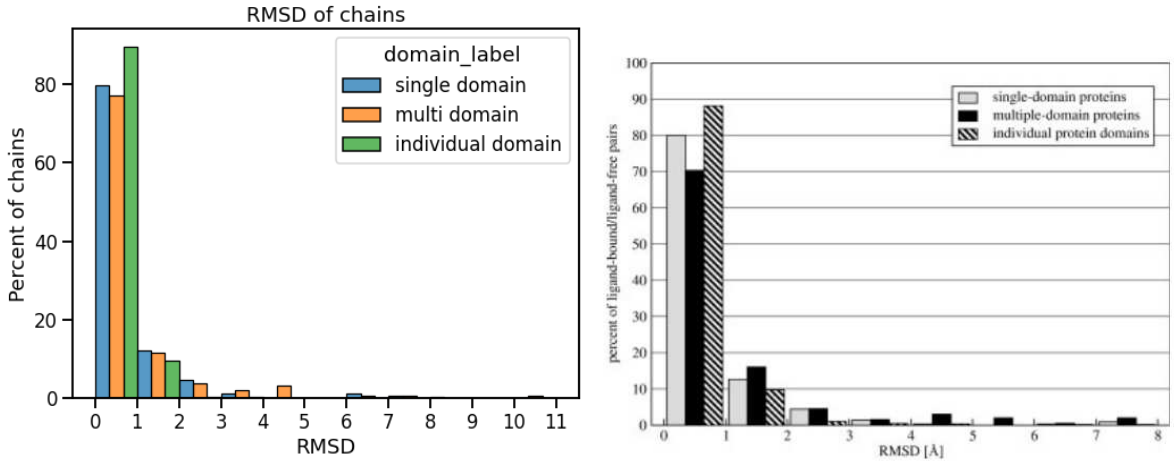


Figure 7. Distribution of the secondary structure similarity across apo and holo form (percentage). The dataset is further subdivided into a dataset of single-domain chains and a dataset of chains where multiple domains were identified (see the label). On the left (ours), the secondary structure is classified into 3 types (helix, sheet, coil), while the plot on the right is based on finer division into 8 types of secondary structure (DSSP), therefore the identity is on average lower.



Ours Theirs

Figure 8. Distribution of the C_alpha RMSD of the apo and holo forms of the entire chains (divided in single-domain or multi-domain chains), and of the individual domains of the multi-domain chains.

Domain movements

Next the focus is on the dataset of two-domain arrangements (as defined in chapter above).

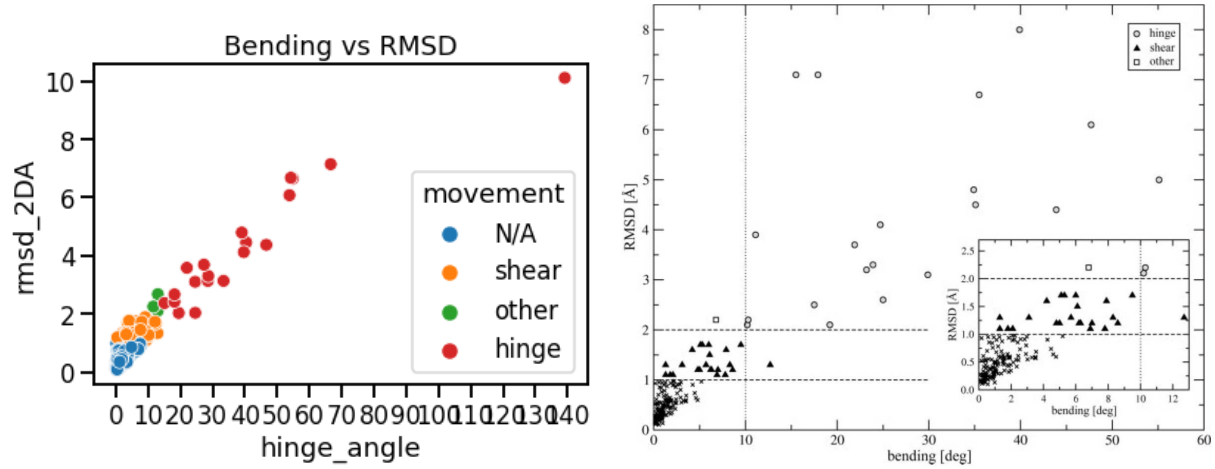


Figure 9. Domain motions upon ligand binding described by the degree of bending and the RMSD between ligand-free and ligand-bound structural forms.

A different method for quantifying the relative domain rotation, hinge angle, is used in (Brylinski and Skolnick, 2007). Our method, described in the preceding chapter, uses superposition to compute the optimal rotation of one of the domains in the pair minimizing the RMSD between its apo and holo form, the angle being computed from the rotation matrix. Whereas their method centers a reference frame (coordinate system) with three axes L, M, S in each of the two domains and computes the angle between the corresponding axes (therefore we have axes L1a, L2a, L1h, L2h, etc, where 1 is first domain, 2 the second and a/h is the structure. We

compute the angle between L1a and L2a; L1h and L2h etc.). This is done for each structure, apo and holo form, and the largest difference of these three angles (|<L1a,L2a> - <L1h, L2h>|, |<M1a, M2a> - <M1h, M2h>|, etc.) in the two forms is chosen as the hinge angle. (To understand why this measurement makes sense, we refer the reader to the paper where they can find the details of how the reference frame is chosen and that the domains are beforehand averaged using the apo and holo form, [using superposition]). The disadvantage is that information about the other two angle differences is lost (while it is possible to describe the rotation completely using one angle). The advantage is that using those axes, the domains can be conveniently visualized (as ellipsoids using these axes and scaling them according to the mass distribution). I am not sure if there is a simple estimate or bound on the differences of values computed using these two methods. In any case, we successfully verified our approach against selected values in a database DynDom (Taylor et al., 2014) which contains a non-redundant set of 1822 domain movements and uses the same method for computing the hinge angle.

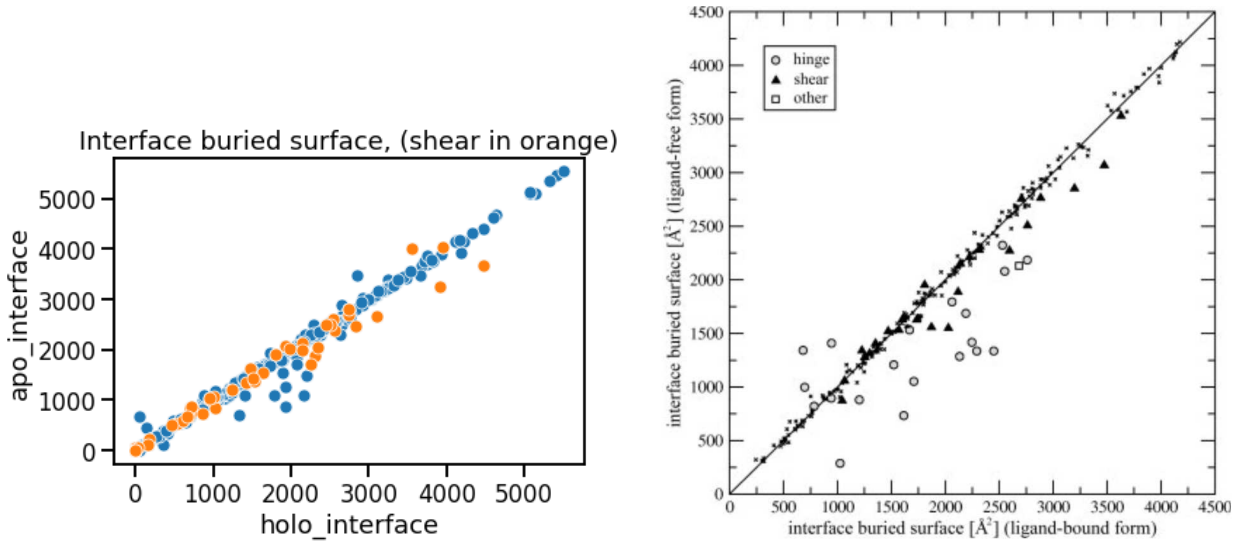


Figure 10. Plot shows the interface area between the two domains before and after ligand binding. Datapoints under the diagonal can be interpreted as the domain closure upon ligand binding.

The paper includes a table with results for selected proteins. The results were in most cases comparable (see attached tables, theirs and ours, in Supplementary material). There are however a few large differences in measured domain interface area. That can be explained by different domain boundary definitions, as illustrated in the following table and figure.

Protein	PDB code, chain: Apo	PDB code, chain: Holo	Two-domain arrangement[a]	Interface buried surface area [Å²]: Apo	Interface buried surface area [Å²]: Holo	Large-scale movement: Bending [deg]	Large-scale movement: RMSD [Å]	Large-scale movement: Type[b]
DAHPS[c]	1vr6A	1rzmA	D1 (1–64) D2 (65–338)	288.0	1024.7	39.9	8.0	[D-h-2]

Apo chain	Holo chain	Two-domain arrangement	Apo interface	Holo interface
1vr6A	1rzmA	D1 (1-70) D2 (71-338)	695	1325

Table 1. The upper table shows measurements for a two-domain arrangement, where there is a significant difference in the measured domain interface area between the two domains (in the apo structure). This can be explained by the different domain definitions. See Figure 11. below for illustration.

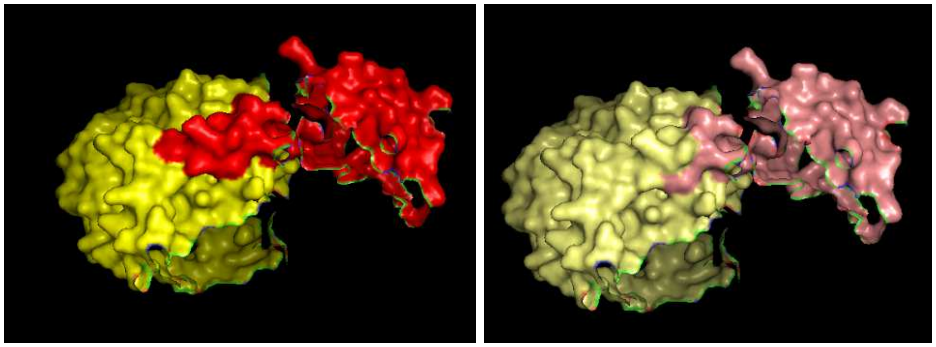


Figure 11. Two-domain arrangement in apo structure 1vr6A. On the left domain boundaries, visualized on the surface of the protein, of our results (colors show distinct domains), on the right domain boundaries of (Brylinski and Skolnick, 2007). This difference explains the larger measured buried interface area between the two domains in our results.

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 comparison of the results on the current PDB.

5.3 Results on recent dataset

In order to fulfill the aim of the work we run the pipeline on a new dataset which includes the structures resolved since the previous work of (Brylinski and Skolnick, 2007).

The dataset (as of 04/29/2022) consists of 595,878 chains, of 170,634 unique structures from PDB. For interpreting the results, we selected one pair per unique Uniprot primary accession, removing the redundancy, resulting in 4674 chain pairs (see the results in chapter 5.3).

Due to the large size of the dataset we run the pipeline on grid infrastructure operated by MetaCentrum. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Some plots based on the *entire current* PDB could look more similar to the plots from the paper, than the plots representing the paper's dataset (Figure 12). As they do not suffer from the skewing due to the important parts of the *non-redundant* dataset missing, e.g. due to the different ligand binding state classification (see previous chapter).

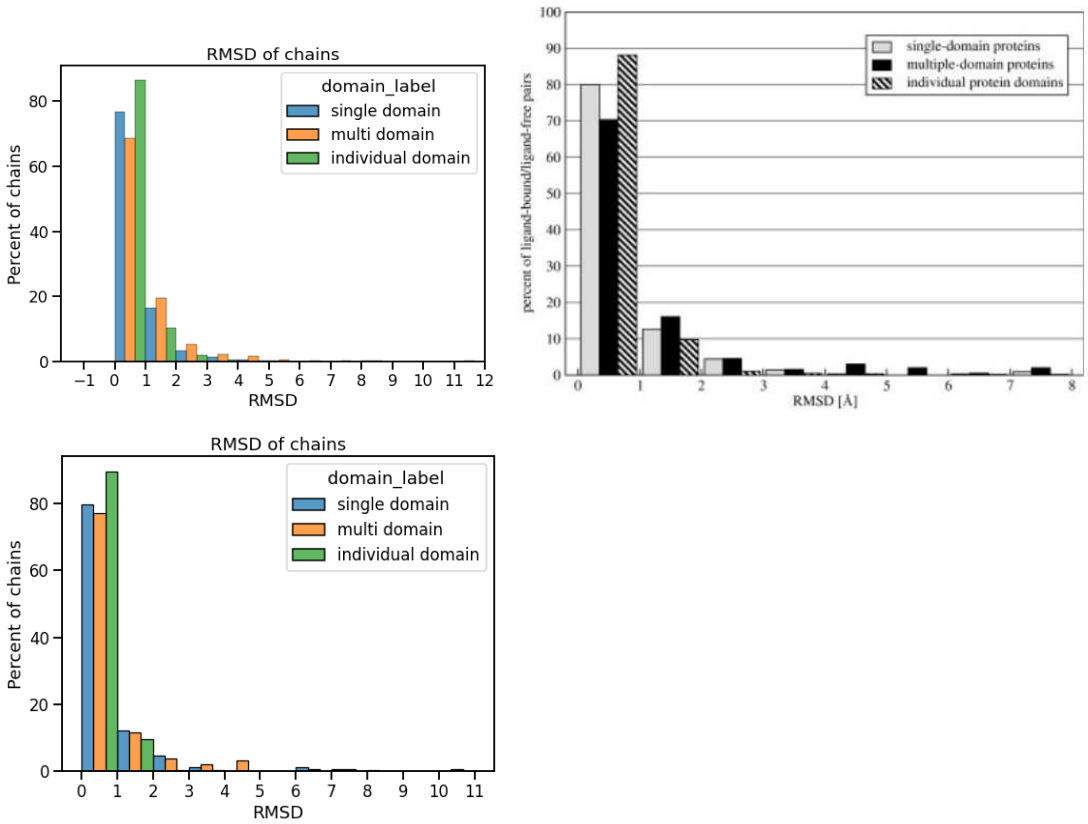


Figure 12. Distribution of the C_alpha RMSD of the apo and holo forms of the entire chains (divided in single-domain or multi-domain chains), and of the individual domains of the multi-domain chains. Top left is the current PDB, top right the previous work and bottom is the previous work dataset (the same plot as in previous chapter).

We observe there is a clear relationship between the magnitude of change in domain interface area and the (apo or holo) interface buried surface (Figure 13). The larger the interface between domains, the lower the change of the interface due to ligand binding.

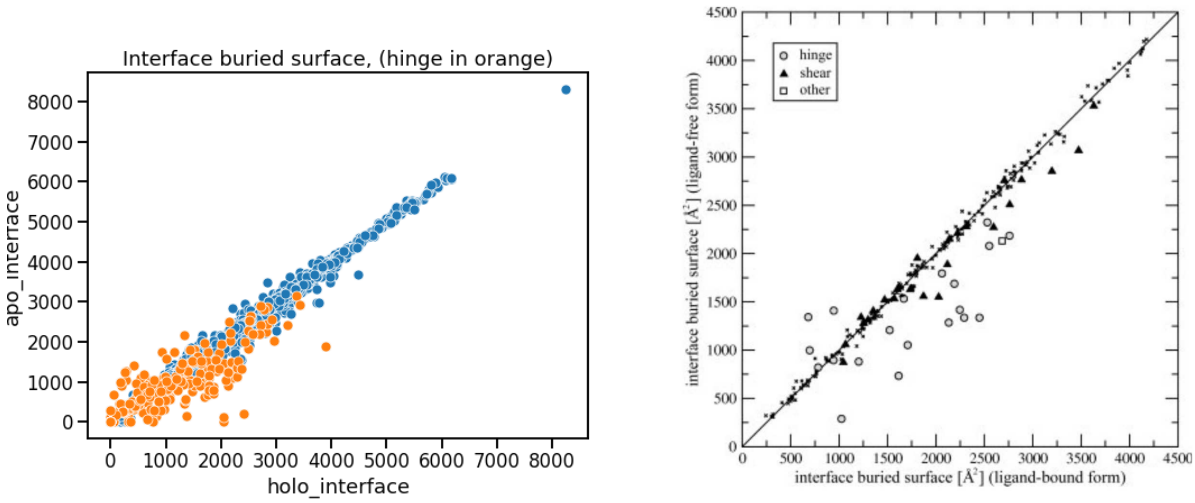


Figure 13. Plot shows the interface area between the two domains before and after ligand binding. Datapoints under the diagonal can be interpreted as the domain closure upon ligand binding. On the right, the figure of (Brylinski and Skolnick, 2007), based on a PDB of (October 2006).

6 Conclusion

We developed a software framework for analyzing apo and holo forms of protein structures. We successfully verified it against previous work and showed it can be run on the current version of PDB.

Although this work did not formulate hypotheses for which we would collect data and discuss the results, and the results served as validation of the method, we can already show the richer dataset since the previous work, the dataset of current PDB, may be used to discover more patterns in the data.

7 References

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

Ahmad, E., Rabbani, G., Zaidi, N., Khan, M.A., Qadeer, A., Ishtikhar, M., Singh, S., Khan, R.H., 2013. Revisiting ligand-induced conformational changes in proteins: essence, advancements, implications and future challenges. J. Biomol. Struct. Dyn. 31, 630–648. https://doi.org/10.1080/07391102.2012.706081

Alexandrov, N., Shindyalov, I., 2003. PDP: protein domain parser. Bioinformatics 19, 429–430. https://doi.org/10.1093/bioinformatics/btg006

Atkins, P.W., De Paula, J., 2006. Atkins' Physical chemistry. W.H. Freeman, New York.

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
CATH [WWW Document], n.d. URL https://www.cathdb.info/wiki (accessed 7.19.22).
Cheng, Y., Grigorieff, N., Penczek, P.A., Walz, T., 2015. A Primer to Single-Particle Cryo-Electron Microscopy. Cell 161, 438–449. https://doi.org/10.1016/j.cell.2015.03.050
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
Einstein, A., 1905. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann. Phys. 322, 549–560. https://doi.org/10.1002/andp.19053220806
Gerstein, M., Lesk, A.M., Chothia, C., 1994. Structural Mechanisms for Domain Movements in Proteins. Biochemistry 33, 6739–6749. https://doi.org/10.1021/bi00188a001
Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J.M., Orengo, C.A., 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35, D291–D297. https://doi.org/10.1093/nar/gkl959
Greener, J.G., Filippis, I., Sternberg, M.J.E., 2017. Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. Struct. England1993 25, 546–558. https://doi.org/10.1016/j.str.2017.01.008
Hayward, S., 1999. Structural principles governing domain motions in proteins. Proteins Struct. Funct. Bioinforma. 36, 425–435. https://doi.org/10.1002/(SICI)1097-0134(19990901)36:4<425::AID-PROT6>3.0.CO;2-S
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).
Hu, Y., Cheng, K., He, L., Zhang, X., Jiang, B., Jiang, L., Li, C., Wang, G., Yang, Y., Liu, M., 2021. NMR-Based Methods for Protein Analysis. Anal. Chem. 93, 1866–1879. https://doi.org/10.1021/acs.analchem.0c03830
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., 1976. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 32, 922–923. https://doi.org/10.1107/S0567739476001873
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
Keedy, D.A., Fraser, J.S., van den Bedem, H., 2015. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. PLoS Comput. Biol. 11, e1004507. https://doi.org/10.1371/journal.pcbi.1004507
Keskin, O., 2007. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. BMC Struct. Biol. 7, 31. https://doi.org/10.1186/1472-6807-7-31
Kromann, J.C., 2022. Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation.

Lovell, S.C., Davis, I.W., Arendall III, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., 2003. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins Struct. Funct. Bioinforma. 50, 437–450. https://doi.org/10.1002/prot.10286
Marion, D., 2013. An Introduction to Biological NMR Spectroscopy. Mol. Cell. Proteomics MCP 12, 3006–3025. https://doi.org/10.1074/mcp.O113.030239
McPherson, A., Gavira, J.A., 2013. Introduction to protein crystallization. Acta Crystallogr. Sect. F Struct. Biol. Commun. 70, 2–20. https://doi.org/10.1107/S2053230X13033141
Mitternacht, S., 2016. FreeSASA: An open source C library for solvent accessible surface area calculations. https://doi.org/10.12688/f1000research.7931.1
Nusrat, S., Khan, R.H., 2018. Exploration of ligand-induced protein conformational alteration, aggregate formation, and its inhibition: A biophysical insight. Prep. Biochem. Biotechnol. 48, 43–56. https://doi.org/10.1080/10826068.2017.1387561
PDBe-KB consortium, 2022. PDBe-KB: collaboratively defining the biological context of structural data. Nucleic Acids Res. 50, D534–D542. https://doi.org/10.1093/nar/gkab988
Pinzi, L., Rastelli, G., 2019. Molecular Docking: Shifting Paradigms in Drug Discovery. Int. J. Mol. Sci. 20, 4331. https://doi.org/10.3390/ijms20184331
Shafee, T., 2017. English: Hydrogen bonds in protein secondary structure. Cartoon above, atoms below with nitrogen in blue, oxygen in red (PDB: 1AXC).
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
Taylor, D., Cawley, G., Hayward, S., 2014. Quantitative method for the assignment of hinge and shear mechanism in protein domain movements. Bioinformatics 30, 3189–3196. https://doi.org/10.1093/bioinformatics/btu506
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
Ulrich Nienhaus, G., Müller, J.D., McMahon, B.H., Frauenfelder, H., 1997. Exploring the conformational energy landscape of proteins. Phys. Nonlinear Phenom., 16th Annual International Conference of the Center for Nonlinear Studies 107, 297–311. https://doi.org/10.1016/S0167-2789(97)00097-3
Wiedemann, C., Kumar, A., Lang, A., Ohlenschläger, O., 2020. Cysteines and Disulfide Bonds as Structure-Forming Units: Insights From Different Domains of Life and the Potential for Characterization by NMR. Front. Chem. 8.
Zlir'a, 2013. Українська: Фрагмент пептидного ланцюга. Кожна пептидна група має планарну структуру, обертання можливе тільки навколо зв'язків C—Cα і N—Cα.

Supplementary material

Contains all the results for proteins and two domain arrangements available in the paper (Brylinski and Skolnick, 2007). The second table is our results.

PDB code, chain			Interface buried surface area [Å^2]		Large-scale movement		
Apo	Holo	Two-domain arrangement[a]	Apo	Holo	Bending [deg]	RMSD [Å]	Type[b]
1vr6A	1rzmA	D1 (1–64) D2 (65–338)	288.0	1024.7	39.9	8.0	[D-h-2]
1usgA	1usiA	D1 (1–124, 247–333) D2 (125–246, 334–345)	1336.1	2291.9	17.9	7.1	[D-h-2]
4akeA	1akeA	D1 (1–121, 160–214) D2 (122–159)	733.3	1614.8	15.5	7.1	[D-h-2]
1cb6A	1bka_	D2 (91–250, 329–341) D3 (342–434, 595–691)	1342.4	681.3	35.5	6.7	[D-h-2]
		D1 (4–90, 251–328) D2 (91–250, 329–341)	1417.3	2246.0	47.7	6.1	[D-h-2]
		D1 (4–90, 251–328) D3 (342–434, 595–691)	1313.4	1298.9	5.7	1.3	[D-s-2]
1sw5A	1sw2A	D1 (6–110, 21–275) D2 (111–210)	1687.5	2190.2	55.1	5.0	[D-h-2]
1y3qA	1y3nA	D1 (1–135, 308–399) D2 (136–307, 400–490)	2084.6	2759.8	34.9	4.8	[D-h-2]
1gudA	1rpjA	D1 (1–112, 247–288) D2 (113–246)	1285.0	2132.2	35.1	4.5	[D-h-2]
1ex6B	1ex7A	D1 (1–32, 84–186) D2 (33–83)	880.5	1202.9	43.9	4.4	[D-h-2]
1wd7B	1wcwA	D1 (8–40, 166–261) D2 (41–165)	1408.3	945.2	24.7	4.1	[D-h-2]
1w0jE	1e1rF	D1 (1–73) D2 (74–466)	1532.5	1667.9	11.1	3.9	[D-h-2]
1rf5A	1rf4A	D1 (1–19, 230–427) D2 (20–229)	1336.1	2449.1	21.9	3.7	[D-h-2]
1s2oA	1tj5A	D1 (1–83, 162–244) D2 (84–161)	1207.2	1522.5	23.9	3.3	[D-h-2]
1k5hA	1q0qA	D1 (1–147) D2 (323–398)	895.6	943.1	23.2	3.2	[D-h-2]
1zolA	1o03A	D1 (1–16, 84–221) D2 (17–83)	1052.5	1708.1	29.9	3.1	[D-h-2]
1viyC	1vhlA	D1 (0–32, 96–207) D2 (33–95)	820.9	783.7	25.0	2.6	[D-h-2]
1gqzA	2gkeA	D1 (1–116, 261–274) D2 (117–260)	1794.4	2062.5	17.5	2.5	[D-h-2]
1za1A	1q95A	D1 (1–136, 292–310) D2 (137–291)	2322.4	2530.2	10.3	2.2	[D-h-2]
1tjdA	1eejB	D1 (1–66) D2 (67–216)	997.5	696.4	19.2	2.1	[D-h-2]
1jejA	1jg6A	D1 (1–172, 335–351) D2 (173–334)	2079.8	2550.6	10.2	2.1	[D-h-2]
1hooB	1cg0A	D1 (1–100, 202–431) D2 (101–201)	2131.2	2683.4	6.8	2.2	[D-?-2]
1e5lA	1e5qA	D2 (130–258, 340–397, 437–450) D3 (259–339)	2271.8	2595.5	9.5	1.7	[D-s-2]
		D1 (2–129, 398–436) D2 (130–258, 340–397, 437–450)	2852.3	3197.6	7.9	1.6	[D-s-2]
1hw1B	1h9gA	D1 (5–82) D2 (83–227)	872.8	1043.8	6.0	1.7	[D-s-2]
1l0wB	1g51A	D3 (142–240, 526–549) D4 (241–294, 414–525)	3528.7	3629.7	5.3	1.7	[D-s-2]
		D5 (295–413) D4 (241–294, 414–525)	1339.9	1226.5	3.1	1.3	[D-s-2]
		D2 (106–141, 550–562) D3 (142–240, 526–549)	1531.9	1569.8	1.3	1.3	[D-s-2]
1otjD	1gy9A	D1 (4–129, 235–282) D2 (130–234)	3068.5	3473.7	5.1	1.7	[D-s-2]

PDB code, chain			Interface buried surface area [Å^2]		Large-scale movement		
Apo	Holo	Two-domain arrangement[a]	Apo	Holo	Bending [deg]	RMSD [Å]	Type[b]
1evkA	1evlA	D1 (242–531) D2 (532–642)	1628.8	1735.0	4.2	1.6	[D-s-2]
1g6wD	1k0bC	D2 (264–298) D3 (189–263, 299–341)	1058.5	1066.0	6.1	1.5	[D-s-2]
1njgB	1njfA	D1 (5–177) D2 (178–243)	1520.5	1470.3	12.7	1.3	[D-s-2]
1t6kA	1u1wA	D1 (1–117, 265–278) D2 (118–264)	1550.1	2028.1	8.3	1.3	[D-s-2]
1yl5B	1yl7A	D1 (1–106, 211–245) D2 (107–210)	1279.1	1253.3	8.6	1.2	[D-s-2]
2cgkB	2cgjA	D2 (76–237) D3 (238–300, 375–480)	2496.6	2761.4	7.8	1.1	[D-s-2]
		D1 (2–75) D3 (238–300, 375–480)	1557.9	1871.7	7.0	1.2	[D-s-2]
1rkaA	1gqtA	D1 (4–163, 242–308) D2 (164–241)	2278.7	2314.5	6.3	1.2	[D-s-2]
1wxdB	2cy0A	D1 (1–103, 231–262) D2 (104–230)	1886.0	2118.3	6.2	1.2	[D-s-2]
2a5aA	1uk4A	D1 (3–197) D2 (198–303)	1628.7	1601.5	5.0	1.2	[D-s-2]
2g26A	2fs4B	D1 (3–15, 142–333) D2 (16–141)	2757.4	2706.8	4.8	1.2	[D-s-2]
1mmiA	1ok7B	D2 (122–247) D3 (248–366)	1405.8	1352.4	6.9	1.1	[D-s-2]
1a8d_	1d0hA	D1 (875–1110) D2 (1111–1315)	2763.4	2884.2	2.1	1.1	[D-s-2]
1pdbA	1kmsA	D1 (3–34, 115–186) D2 (35–114)	1951.8	1808.0	1.8	1.1	[D-s-2]
1hfkA	1hg1C	D1 (4–216) D2 (217–327)	2151.9	2140.4	1.8	1.1	[D-s-2]
1k6wA	1k70A	D1 (4–55, 367–410) D2 (56–366)	2210.9	2224.0	1.3	1.1	[D-s-2]

		apo_interface	holo_interface	rmsd_2DA	spans_auth_seq_id_x	spans_auth_seq_id_y
apo_chain	**holo_chain**					
1vr6A	**1rzmA**	695	1325	10	((1, 70),)	((71, 338),)
1usgA	**1usiA**	885	1920	7	((1, 120), (250, 330))	((121, 249), (331, 345))
1cb6A	**1bka_**	1502	2195	6	((1004, 1091), (1251, 1339))	((1092, 1250),)
		1516	1469	1	((1004, 1091), (1251, 1339))	((1340, 1434), (1595, 1691))
		2503	2294	1	((1340, 1434), (1595, 1691))	((1435, 1594),)
1sw5A	**1sw2A**	1628	2088	5	((6, 109), (214, 275))	((110, 213),)
1y3qA	**1y3nA**	1706	2080	5	((1, 132), (311, 399))	((133, 310), (400, 490))
1gudA	**1rpjA**	1105	1788	4	((1, 110), (250, 278))	((111, 249), (279, 288))
1ex6B	**1ex7A**	1091	1197	4	((233, 292),)	((201, 232), (293, 386))
1wd7B	**1wcwA**	1085	878	4	((8, 42), (165, 261))	((43, 164),)
1rf5A	**1rf4A**	1090	2157	4	((1, 19), (230, 427))	((20, 229),)
1s2oA	**1tj5A**	1108	1397	3	((1, 88),	((89, 159),)

					(160, 244))	
1k5hA	**1q0qA**	824	857	3	((1, 150),)	((312, 398),)
1zolA	**1o03A**	1272	1935	3	((1, 14), (93, 221))	((15, 92),)
1gqzA	**2gkeA**	1557	1898	2	((1, 114), (262, 274))	((115, 139), (150, 261))
1za1A	**1q95A**	2372	2564	2	((1, 133), (292, 310))	((134, 291),)
1jejA	**1jg6A**	2470	2833	2	((1, 166), (333, 349))	((167, 332),)
1hooB	**1cg0A**	1888	2300	2	((1, 100), (201, 265))	((101, 200),)
		3239	3916	2	((1, 100), (201, 265))	((266, 431),)
1e5lA	**1e5qA**	2308	2625	1	((2, 124), (395, 442))	((125, 249), (350, 394), (443, 449))
		1987	2146	2	((125, 249), (350, 394), (443, 449))	((250, 349),)
1hw1B	**1h9gA**	742	857	2	((5, 79),)	((80, 227),)
1l0wB	**1g51A**	217	182	2	((1001, 1106),)	((1137, 1277), (1416, 1580))
		1340	1416	2	((1137, 1277), (1416, 1580))	((1278, 1415),)
1evkA	**1evlA**	1538	1646	2	((242, 530),)	((531, 642),)
1g6wD	**1k0bC**	2021	2056	1	((100, 196),)	((197, 352),)

1njgB	**1njfA**	836	1024	1	((5, 11), (178, 243))	((13, 177),)
1t6kA	**1u1wA**	1716	2248	1	((1, 119), (268, 278))	((120, 267),)
1yl5B	**1yl7A**	1060	1003	1	((1, 105), (213, 245))	((106, 212),)
2cgkB	**2cgjA**	3682	4476	2	((2, 236),)	((237, 480),)
1wxdB	**2cy0A**	1374	1535	1	((1, 99), (236, 249))	((100, 235),)
2a5aA	**1uk4A**	2392	2391	1	((3, 14), (100, 197))	((15, 99),)
		1495	1424	1	((3, 14), (100, 197))	((198, 301),)
2g26A	**2fs4B**	2597	2548	1	((3, 162), (165, 174))	((175, 332),)
1mmiA	**1ok7B**	1710	1666	1	((1, 123),)	((124, 247),)
		1246	1207	1	((124, 247),)	((248, 366),)
1a8d_	**1d0hA**	2498	2637	1	((12, 246),)	((247, 452),)
1hfkA	**1hg1C**	2083	2111	1	((4, 18), (34, 218))	((219, 327),)
1k6wA	**1k70A**	2116	2122	1	((4, 56), (365, 409))	((57, 364),)