ANALYSIS OF MEMBRANE PROTEINS IN THE OPM DATABASE

Undergraduate Project Report

Mentor: Prof. R. Sankararamakrishnan, BSBE

Abhinav Tiwari | 190031 | abtiwari@iitk.ac.in | Biological Sciences and Bioengineering

Indian Institute of Technology, Kanpur 03rd May, 2022

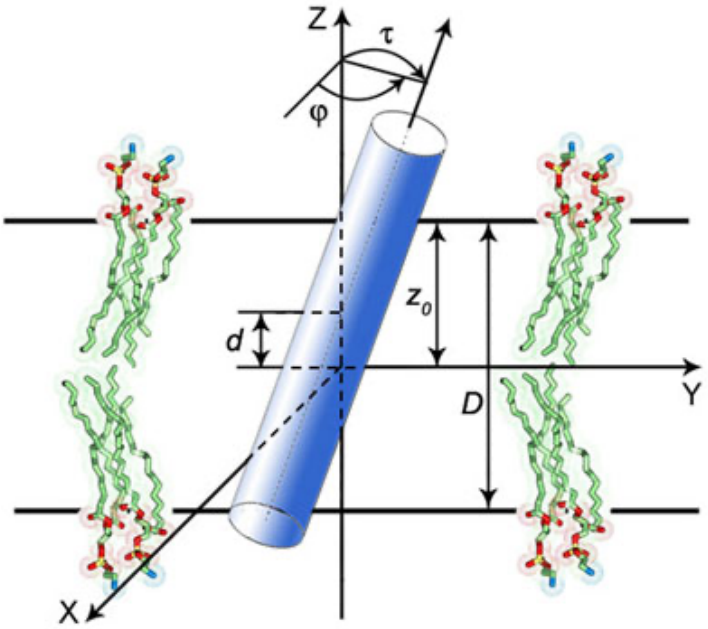
Abstract

Membrane proteins are a very important class of protein which play a very important role in the regular functioning of cells. Since a cell needs to be separated by its surroundings to operate efficiently it still needs to transport substances in and out. Membrane proteins are embedded in the lipid bilayer and provide an interface between the cytoplasm and the periplasm. Some important types of membrane proteins are ion channels, receptors and transporters.A large amounts of drugs nowadays target membrane protein and their receptor binding domains in order to block or regulate their activity. Although the number of known structures of membrane proteins is still small relative to the size of the proteome as a whole, many new membrane protein structures have been determined recently. In this report I will provide a summary on the work I did in analyzing the membrane proteins present in the OPM database which I undertook as my UGP project.

INTRODUCTION

There are many online sources and servers to find information about the structure of proteins. The different databases are designed for different purposes, where some have information specific to membrane proteins such as transporters, receptors etc. In this analysis I have focussed mainly on structural databases such as OPM, mpstruc and PDB. The advantage of using these databases is that many of the protein structures present are already classified and annotated with useful information.

OPM database is a highly curated database that provides relevant information such as position of protein with respect to membranes as well as their structural classification, localization and topology. OPM also has a server which can schedule jobs for calculating the position of protein in the membrane. It works by considering each protein as rigid body that freely floats in a hydrophobic slab of adjustable thickness The orientation of the protein is determined by minimizing an elaborate transfer energy ($\Delta G_{transfer}$) with respect to a shift along the bilayer normal, hydrophobic thickness, rotation angle, and tilt angle in a coordinate system where the membrane is considered normal to the Z axis.

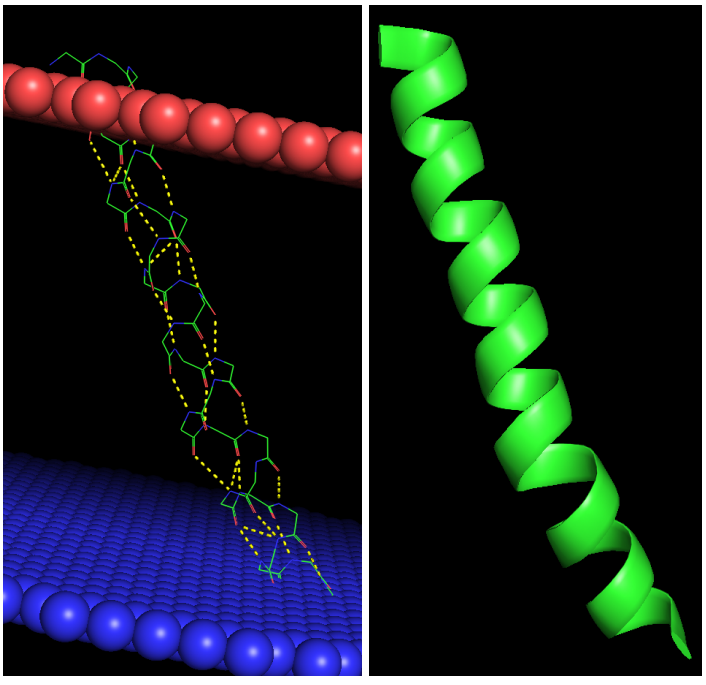


Schematic representation of a transmembrane protein model implemented by OPM server

The PDB database is the most extensive dataset and contains information about the fasta sequence of the proteins which was also used to classify protein structures to remove redundancy using CD HIT .

PROPERTIES OF ALPHA HELICES

Alpha helix is a common secondary structure which is commonly found in the native state of many stable proteins. Alpha proteins are notably favorable for membrane bound proteins due to their geometrical and hydrophobic properties which help them exist in the transmembrane. Alpha helices are generally classified as a stretch of amino acids joined by peptide bonds which form the backbone of the helix. The occurrence of periodic geometrical patterns classify a structure as an alpha helix. The ramachandran angles which measure the psi and phi angles for residues in polypeptides give characteristic values for alpha helices of **φ= -57º and ψ= -47º** There are typically 3.6 residues per turn of the helix repeated periodically with a height increase of 5.4A per turn. Typically the O of **ith** residue interacts with the NH group of the main chain of the **i+4th** residue with polar hydrogen bonds.Because the amino acids connected by each hydrogen bond are four apart in the primary sequence, these main chain hydrogen bonds are called "n to n+4" But the parameters noted above are not constant in practical structure, alpha helices in different proteins vary a lot in terms of their geometric arrangements, these variations give rise to alternate models to study alpha helix structures such as a **3_{10} helix and a pi helix**. The ability to quantify the deformations of flexible elements in a protein fold is paramount for the development of flexible templates in computational *de novo* protein design. The lipid bilayer is generally 5- 7 nm thick and so to cross the bilayer completely a transmembrane segment must be 19-23 residues long, this may also vary depending on the orientation and tilt of the transmembrane segment.



Schematic representation of a 27 residue alpha helix inside the membrane and cartoon structure analyzed using pymol

MATERIALS AND METHODS

Collecting Structures

Protein structures were taken from the OPM database. OPM database classifies proteins into three types namely - transmembrane, peripheral and peptides. Transmembrane proteins are further classified into alpha-helical-polytropic ,Bitopic and beta barrel transmembranes. The OPM Database contains an API which can provide a list of PDB ids of protein from each category. Using this API information for **6568** proteins were obtained out of which **3323** alpha-helical-polytropic proteins were selected for further analysis.In further refinement steps, proteins with more than 1 lipid membranes or proteins having curved membranes were removed for simplicity. The PDB files of these proteins were found from the OPM database which contained the information for the arrangement in membrane and PDB were also obtained from the RCSB PDB Database. RCSB PDB files have α-helix annotation information whereas OPM PDB files have transmembrane region information. When these two pieces of information are brought together, then transmembrane α-helical regions can be properly identified and annotated.

Removing Redundancy

The structure gathered by the discussed process had a lot of redundancy in them, meaning that many proteins were homologous or the same just crystallized with different ligands etc. For our analysis we need just the alpha helices and it should be as diverse as possible, redundant structures may bias our results to a particular type of structures which may be in majority in the dataset. To remove this redundancy the PDB files were first queried with the RCSB database to get the FASTA sequences of all the PDB files. This fasta sequence was queried using the **CD-HIT** tool which is a widely used tool for removing redundancy in protein sequences.Using a sequence similarity cut off of **40%** and other default parameters were used to cluster the sequences. After some rudimentary data cleaning and checking the quality of PDB files, around **663** unique clusters were obtained. One sequence from each cluster was chosen to be a representative sequence for further analysis.

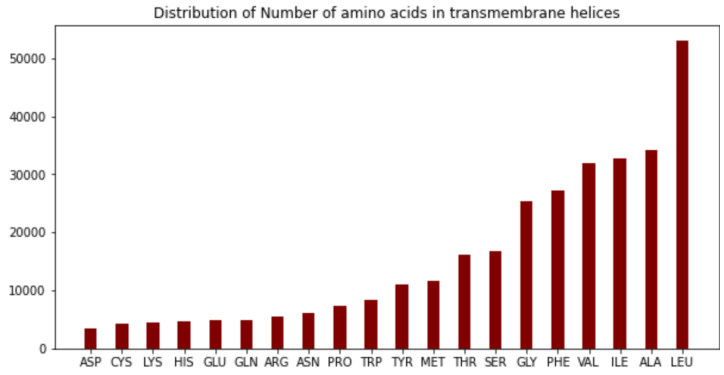
Code Workflow

For all the 663 structures a general workflow was designed to collect the geometrical information out of the sequence and generate reports for further analysis. Using the membrane information obtained using OPM

dataset, membrane boundaries were determined and protein residues were separated on the bases of their presence in or outside the membrane. This was used to create separate PDB files containing just the membrane bound regions. For each structure Helix regions were found based on the information given in the PDB files, and for each helix parameters such as helix length, starting residue, ending residue and chain identifiers were calculated. Then the side chains were removed to analyze the backbone structure of helices. For each helix residue distance between the ith O and i+4th NH was calculated and breaks were characterized in the helix whenever this distance was found to be < 3.5 A. The ramachandran phi and psi angles were also separately calculated and plotted for each structure. In another analysis the whole transmembrane segment of the structures were considered, parameters such as number of transmembrane segment, segment length were found out. Whenever a loop region was found in the transmembrane segment connecting two alpha helices it was characterized as a break in the transmembrane domain(different from breaks in hydrogen bonds). Two separate reports were generated containing the information about the helices and transmembrane segments of all the protein structures in the database.

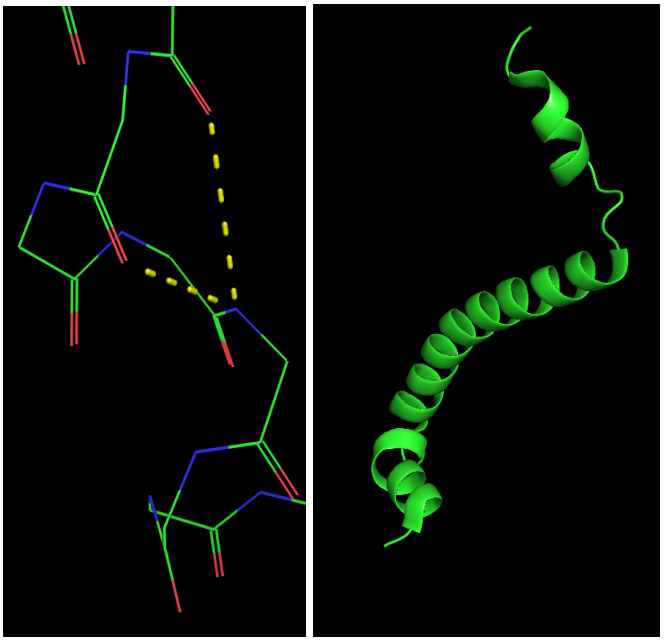
Results and Analysis

Firstly the regions of hydrogen bond break were analyzed and the residues were stored. Firstly all the transmembrane helices were analyzed to know the distribution of amino acids in transmembrane regions of our database. The analysis revealed hydrophobic amino acids such as leucine, alanine valine, isoleucine having the highest presence in our dataset. This was expected because membrane alpha helices are known to have many hydrophobic residues, in order to survive in membrane conditions.



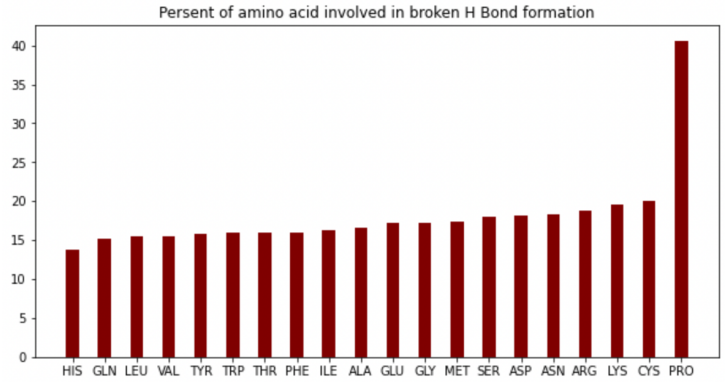
Plot showing the abundance of different residues in our sample

Overall 663 protein structures were analyzed and it was found that 652 structures had at least one broken hydrogen bond in one of its helices. The total number of transmembrane helices analyzed in all the structures were 21180 out of which 14028 helices were found broken which is about **66.23** percent of helices broken.



Example of break in the i , i+4th H bonding rule(left) and presence of loop region between two helices in transmembrane region

For the broken helices, the residues which were responsible for break were identified and their relative abundance calculated. When looking at the absolute values of residues present in borken regions no significant changes from background residue were seen. But when the percent of each residue taking part in the hydrogen bonding was calculated residues proline was found to be present in an broken region more than 40% of time whereas abundant residues such as leucine and alanine were only present 8% of times. This is indeed significant because the background concentration of Leucine is 8 times more than leucine. Next the residues bordering the broken residue segments were analyzed, in this also the same pattern was observed, where Proline had the highest percent of presence in the neighboring residues with 25% followed by Arginine and Lysine. Whereas Leucine had only 19%.



This deformation is due to the Steric crowding between the 5-membered ring of proline residue in the middle of αhelix and the preceding residue that causes a kink the helix.

Since there is a break in at certain places in the alpha helix within the membrane, then it is a good question to ask how such residues are stabilized in the membrane if they cannot form their typical intra main chain hydrogen bond. To find this I checked if the broken residue interacted with any other residue in the helix main chain. Most of the intra-mainchain interactions were found to be with residues i and i + 3, with 17.6% of the residues with broken Hydrogen bond being stabilized by this kind of interaction. Interaction of the form i and i + 5 was also found but was very rare .

Conclusion and Future

The work done so far has described a general procedure for creating a database for membrane proteins with all the essential steps like data collection, cleaning, combining data from different sources, removing redundancy and obtaining a set of representative sequences representing the whole family of transmembrane alpha helical databases. The code also provides a general workflow for analyzing protein geometry and interactions. The analysis and results generated were at most of the time able to replicate several studies already done in the domain.A Study can be done to analyze in more detail how the residues with broken H bonds are stabilized including main chain contacts with other helices, contact with ligands and contact with side chains. A follow up study which has also not been extensively done can be to analyze the loop regions which spawn the transmembrane. I have started working with such analysis but the results are too preliminary to state here. Loops found in the transmembrane regions are of two types, one which is found between membrane, connecting two helices and one at the surface of membranes, which has a U shape, analysis of the kind of residues present in such regions and how they are stabilized in the membrane can be helpful to obtain a overall picture of protein embedded in membrane. There is also scope of machine learning techniques and structural analysis such as PCR to see how the overall structure of the helix in the transmembrane is affected due to the defects described here. I would definitely like to pursue further in the project if given the opportunity and contribute more to this field.

References

- https://opm.phar.umich.edu/about#methods_and_definitions
- https://www.rcsb.org/docs/general-help/membrane-protein-resources
- https://www.sciencedirect.com/science/article/pii/S0005273618300051
- https://www.rcsb.org/downloads/fasta
- http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi
- https://pymol.org/2/
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257318

Data Availability

All the analysis codes and the list of PDB files used in the analysis can be found at the github repository - https://github.com/matrix101A/bioSimulation which will be updated in due time.