Research Article

OPEN ACCESS

Available on http://www.pjbt.org
Pakistan Journal of Biotechnology
(PJBT)
(P-ISSN: 1812-1837 and E-ISSN: 2312-7791)

STRUCTURAL REVELATION OF HUMAN TRANSMEMBRANE PROTEINS BY SVM

Mehwish Faiz[1,] Fariha Ibrahim[1], Shahzad Nasim[3], Choudhary Sobhan Shakeel[1] and Bullo Saifullah[2*]

[1]Ziauddin University (FESTM), Department of Biomedical Engineering, Karachi, 74200, Pakistan
[2]Department of Human and Rehabilitation Sciences, Begum Nusrat Bhutto Women University Sukkur, Sindh, 5200.
[3]Department of Management Sciences & Technology, Begum Nusrat Bhutto Women University Sukkur, Sindh, 65200.

*Corresponding author email: saifullah.bullo@bnbwu.edu.pk

Article Received 18-11-2022, Article Revised 11-01-2023, Article Accepted 27-02-2023.

ABSTRACT

Proteomics is a multidisciplinary research area with proteins as the key element. One of the main constituents of the architecture of the human body is a transmembrane protein with an amino acid as its individual unit at the micro-level. These amino acid sequences yield structural information by unfolding the orientation in 3D space and are critical for the underlying diseases. Moreover, with the rise of the number of proteins in data banks, machine learning algorithms are a savior to reveal this information in no time. We implemented a Support Vector Machine (SVM) on the Protein Data Bank of Transmembrane protein (PDBTM) to harvest the secondary structure of the protein as alpha-helices. The key feature of our approach is that the graphical user interface shows the intensity of the helices in a protein by the amount of spirals as a percentage. Higher values reveal more spirals at the secondary structure level and vice versa.

Keywords: *Transmembrane Protein (TMP), Support vector machine (SVM), Secondary Structure, Alpha Helices (AH).*

INTRODUCTION

Protein is an imperative biomolecule in humans responsible for maintaining vital functionalities including enzymatic activities, hormonal changes, transport mechanisms, and providing defense against diseases (Murray et al., 2017). The main subunit of this biomolecule is an amino acid. More than 300 amino acids have been discovered, and only 20 amino acids participate in protein synthesis (Akram et al., 2011). These amino acids while forming protein, are folded because of the various intermolecular and intramolecular forces to yield a complex 3-D orientation. The simplest structure of a protein is a linear sequence of amino acids named primary structure. This linear chain is assembled via hydrogen bonding and formed as alpha helices or beta-sheets and is termed as secondary structure. This secondary structure tends to form a tertiary structure which when again folded, gives rise to the most complex level of the organization referred to as the quaternary structure with a prosthetic group. Among these different structures, the most functional is the secondary structure that leads to identifying the further complex structural level of protein (Murray et al., 2017) The membrane protein is a type of protein that comprises 30% of sequenced genes. These membrane proteins are present at the cell membrane or at the membrane of the organelles of the cell of human beings. Since they are located at the membrane, they specifically act as a gateway for the transmission of signals and materials (Landreh & Robinson, 2014), (Congreve & Marshall, 2010).

Membrane proteins are classified mainly into two types:

- Transmembrane Alpha helices
- Transmembrane beta-barrel

Figure 1. Depicts the structural orientation of the secondary structure of a protein

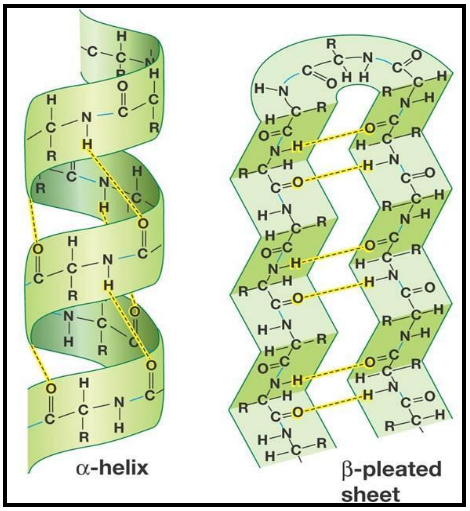


Figure 1. Alpha Helical & Beta sheet structure (retrieved from https://www.chegg.com/flashcards/cell-biology-final-f3d5df4f-e599-4f81-882a-c7fe8ceb24d7/deck)

Transmembrane alpha-helices are the most occurring secondary structure-based membrane proteins (Heijne, 1996). These proteins are involved in an extensive range of important biological processes including the transport of membrane-impermeable molecules and cell-cell communication. Experimental prediction of protein structures and functions in laboratories is quite difficult. Moreover, these methods can predict a few proteins at a given time and cannot handle thousands of proteins at a time. Thus, computational tools are designed to counteract these drawbacks (Cai et al., 2006). The structural information of a protein is critical as structure directly relates to the function of a specific protein. The slight difference in the orientation of protein may lead to the misfolding of the protein chain which ultimately leads to mutation. Therefore, the predictive machine learning algorithms reveals the misfolding/mutation in a protein chain which help to determine related disorder, dysfunctions, and diseases (Congreve & Marshall, 2010), (Conn et al., 2007). Accessing the annotation, topology, and orientation of any protein could be possible through structural information and knowledge of structure can be helpful in relation to drug designing, genomic chip designing as well as artificial cell formation. It can be helpful for clinicians to diagnose the causes of diseases related to protein such as mutations, metabolic diseases, and neuro-generative diseases, and can also determine causes of protein degradation and instability (Tsai et al., 2000).

Machine learning techniques make use of amino acid sequences of a protein as an input to yield the secondary structure information. It is an intermediate but useful step for the prediction of three-dimensional (3D) protein as well as the complex 3-D structure as still it cannot be accurately revealed directly from the sequences (Koch & Schäfer, 2018).

The most critical structural level where the folding of a linear chain of amino acids occurs is the secondary structure. The main element behind this twisting is the involvement of intermolecular forces which defines the basic orientation of the protein. This orientation yields not only information regarding the quaternary structure but also about the functionality. There are various experimental methods that affirm the secondary structure of proteins including X-ray crystallography, nuclear magnetic resonance (NMR) (Dokholyan, 2020) , and electron microscopy (Apweiler, 2001). However, each method has its constraints as they are unable to identify the transient or stable complexes in a cell mostly. Another drawback is that they need a huge and expensive experimental setup, which may take months or even years for the results of the procedure. Due to these limitations of experimental approaches for secondary structure prediction of protein, computational approaches are used. The first computational approach used for forecasting the secondary structure of globular proteins is the Chou-Fasman algorithm. This method is based on assigning a value to each amino acid on their recurrence in the alpha-helical strand or beta-sheet into six different groups. The main drawback of this approach is the low accuracy and unreliable parameters which leads to over-prediction. With this approach, the secondary structure of the protein can be predicted with 50-60% *accuracy* (Kubota et al., 2014).

Another approach relating to the Garnier–Osguthorpe–Robson (GOR) Algorithm has also been employed for the prediction of the secondary structure of a protein. The GOR approach considers not only the probability of each amino acid for the specific secondary structure but also the conditional probability of each monomer of the protein chain. This algorithm yields an accuracy of 65% because it frequently predicts the beta-sheets as loops or chaotic regions (singh et al., 2013). A recent development is the use of the PSI-BLAST, a program used to widen the profile by including more evolutionarily related matches and it results in increased accuracy (Tenreiro Machado et al., 2013).

The Nearest Neighbor Method works on the principle of neighboring values in the vicinity. In proteomics, it uses the sequence similarity of known proteins of the secondary structure of the target protein through a sliding window. It makes use of the available various similarity matches of known structures. The two well-known servers based on the nearest neighbor are NNSSP and Praetor. These two approaches make use of the sequence alignment approach, however, they are distinct as one approach utilizes pair-wise alignment and the other adopted multiple sequence alignments (Conn et al., 2007), (Rask-Andersen et al., 2011).

The Hidden Markov Model is a statistical approach in order to forecasting the secondary structure of the protein. It makes use of small segments of similar sequences of amino acids with known structures to create multiple sequence alignment profiles to generate the hidden Markov Models (HMM) which ultimately infer the structure of the unknown protein (Shukla et al., 2012). Based on this HMM approach, Bystroff et al. developed a program HMMSTR exhibiting an accuracy of 74.3% (Das et al., 2015).

Using NNs, the methodology (Barve et al., 2013) adopted three-layer feed-forward NNs with the inclusion of evolutionary information using multiple sequence alignments. And it showed outstanding performance of Q3=70.8% on 126 non-homologous data sets (RS126). Besides this approach, there are many other approaches using different NN architectures.

A recent SVM-based approach makes use of frequency profiles with evolutionary information as an encoding scheme for the structural analysis of proteins at the secondary level of organization. This approach is applied to the CB513 dataset with the accuracy of Q3=*73.5* (Cong et al., 2013).

Another SVM based inherited two layers of SVM with a weighted cost function for balanced training and it presented a prediction accuracy of Q3=71.5 on the C396 set. Also, there was another scheme that incorporated PSI-BLAST Position Specific Scoring Matrix (PSSM) profiles as an input vector and that applied new tertiary classifiers. This scheme, which is called SVM psi, showed the prediction accuracy of Q3=76.6 on the CB513 data set (Landreh & Robinson, 2014).

A study was carried out for the prediction of alpha-helical transmembrane proteins with the aid of a deep transfer learning technique. Data from two protein classes including alpha-helical polytopic proteins and biotopic proteins were acquired from a dataset referred to as orientations of proteins in membranes. The results of the study exhibited that the proposed technique exhibited good classification accuracy (Wang et al., 2022).

In a similar study, multiscale deep learning fusion was executed for the prediction of alpha-helical transmembrane proteins. The technique involved two modules. The first module includes a prediction of the transmembrane helix through tail modeling. The second module comprises orientation modeling achieved by the SVM classifier. The results of the study demonstrated a reliable classification of transmembrane proteins (Feng et al., 2020).

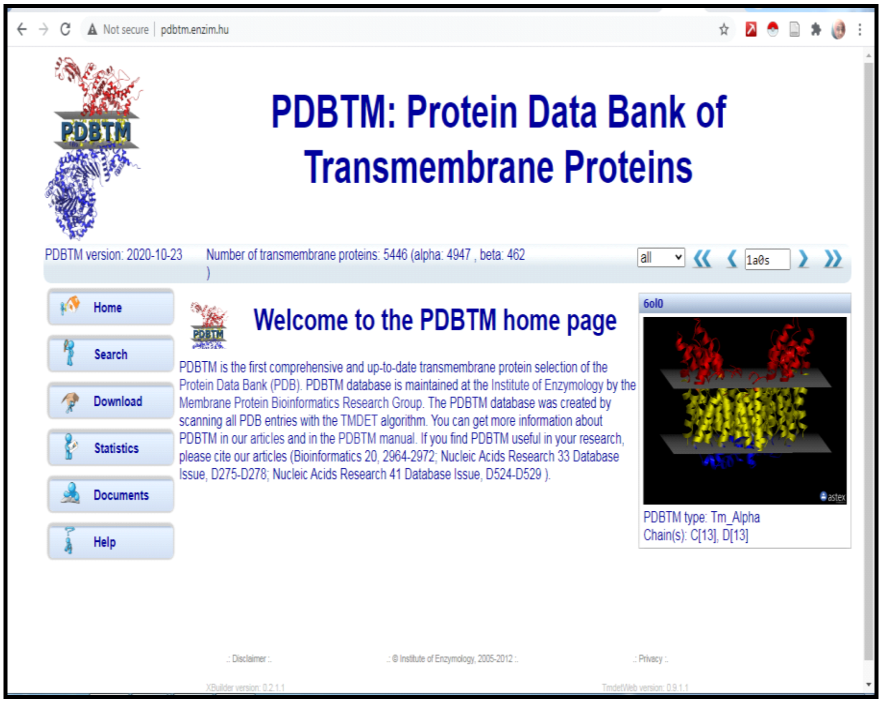
Prediction of interaction sites in alpha-helical transmembrane proteins was carried out utilizing the deep learning method. A stacked ensemble technique was fused with deep learning residual neural networks and the proposed framework was able to achieve an accuracy of 68.9% (Sun & Frishman, 2021).

In a similar study, structural features of transmembrane proteins were classified. Data was acquired from orientations of proteins in the membranes database. A deep learning algorithm consisting of neural networks was applied in addition to a random forest machine learning classifier. The results of the study showed that the proposed prediction technique was reliable with an accuracy of 70% being achieved (Hönigschmid et al., 2020).

From the literature, it is evident that the best computational technique for analyzing protein is either Hidden Markov Model or SVM in terms of accuracy. Moreover, there are very few methods for inferring the secondary structure of the transmembrane protein but our proposed method is unique as it is also revealing the number of alpha-helices present in a transmembrane protein through a Graphical User Interface.

MATERIALS AND METHODS

Dataset: For transmembrane proteins, we retrieved data from the Protein Data Bank of Transmembrane Protein database (http://pdbtm.enzim.hu) which is the first extensive database for transmembrane proteins. This database is unceasingly updated on regular basis by the TMDET algorithm. From PDBTM, we extracted the alpha-helical proteins only, which were around 3729 proteins. The file from PDBTM contains the amino acid sequence only in FASTA Format, with no information to which organism these proteins belong. Figure 2 demonstrates the official website of PDBTM.



PDBTM: Protein Data Bank of Transmembrane Proteins

PDBTM version: 2020-10-23 Number of transmembrane proteins: 5446 (alpha: 4947 , beta: 462)

Welcome to the PDBTM home page

PDBTM is the first comprehensive and up-to-date transmembrane protein selection of the Protein Data Bank (PDB). PDBTM database is maintained at the Institute of Enzymology by the Membrane Protein Bioinformatics Research Group. The PDBTM database was created by scanning all PDB entries with the TMDET algorithm. You can get more information about PDBTM in our articles and in the PDBTM manual. If you find PDBTM useful in your research, please cite our articles (Bioinformatics 20, 2964-2972; Nucleic Acids Research 33 Database Issue, D275-D278; Nucleic Acids Research 41 Database Issue, D524-D529).

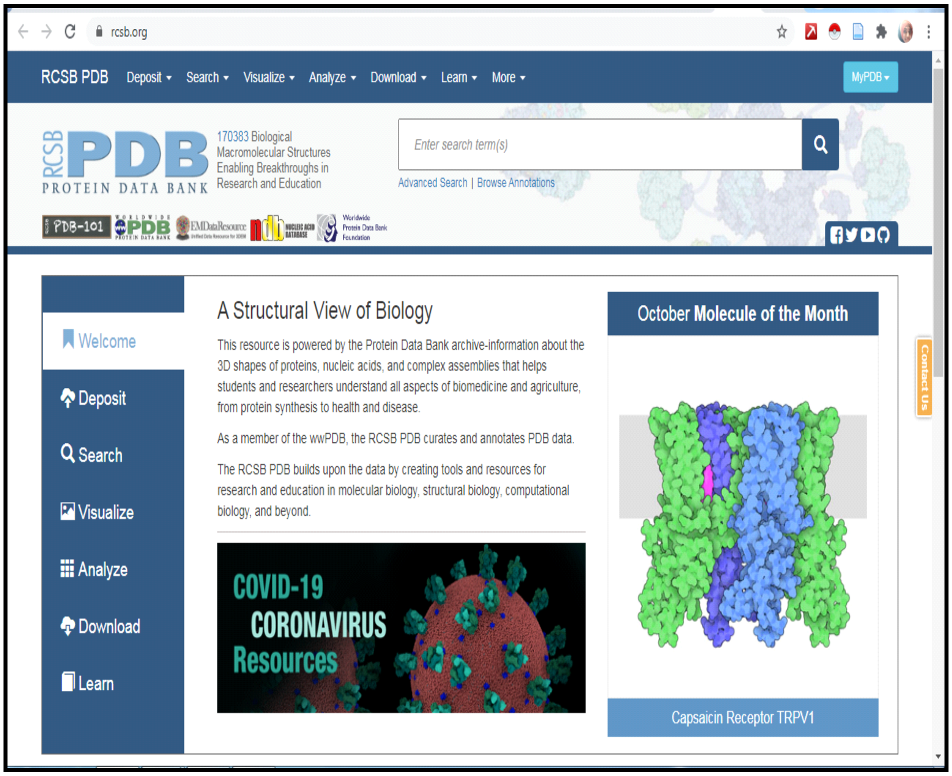
PDBTM type: Tm_Alpha
Chain(s): C[13], D[13]

.: Disclaimer :. .: © Institute of Enzymology, 2005-2012 :. .: Privacy :.

XBuilder version: 0.2.1.1 TmdetWeb version: 0.9.1.1

Figure 2. PDBTM, the protein data base

In order to extract this information about each protein, RCSB PDB (https://www.rcsb.org/) is used. Every protein is sorted and assigned to the species to which it belongs. These proteins belong to various organisms including Bos taurus, Thermosynechococcus vulcanus, Squalus acanthias,Halobacterium salinarum, human beings, and many other organisms. Figure 3 depicts the web page of RCBS PDB.


rcsb.org
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
Enter search term(s)
Advanced Search | Browse Annotations
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataResource
ndb NUCLEIC ACID DATABASE
Worldwide Protein Data Bank Foundation
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Contact Us
COVID-19 CORONAVIRUS Resources

170383 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

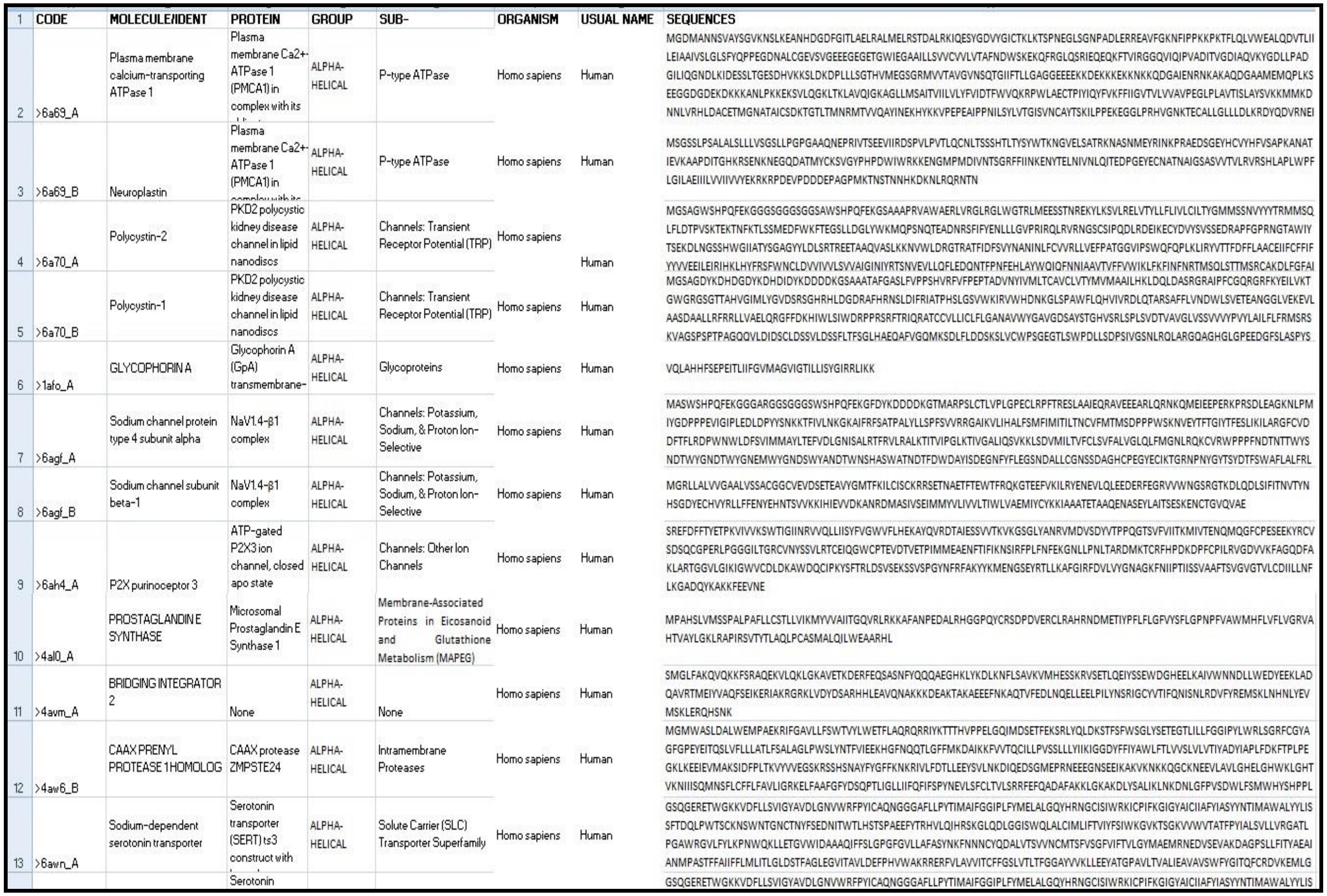
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

October Molecule of the Month

Capsaicin Receptor TRPV1

Figure 3. Web page of RCBS PDB

Further scrutinizing of the data is done to obtain the human proteins only. The total number of human proteins obtained/downloaded is 419 which is shown in figure no.4.



1	CODE	MOLECULE/IDENT	PROTEIN	GROUP	SUB-	ORGANISM	USUAL NAME	SEQUENCES
2	>6a69_A	Plasma membrane calcium-transporting ATPase 1	Plasma membrane Ca2+-ATPase 1 (PMCA1) in complex with its	ALPHA-HELICAL	P-type ATPase	Homo sapiens	Human	MGDMANNSVAYSGVKNSLKEANHDGDFGITLAELRALMELRSTDALRKIQESYGDVYGICTKLKTSPNEGLSGNPADLERREAVFGKNFIPPKKPKTFLQLVWEALQDVTLIILEIAAIVSLGLSFYQPPEGDNALCGEVSVGEEEGEGETGWIEGAAILLSVVCVVLVTAFNDWSKEKQFRGLQSRIEQEQKFTVIRGGQVIQIPVADITVGDIAQVKYGDLLPADGILIQGNDLKIDESSLTGESDHVKKSLDKDPLLLSGTHVMEGSGRMVVTAVGVNSQTGIIFTLLGAGGEEEEKKDEKKKEKKNKKQDGAIENRNKAKAQDGAAMEMQPLKSEEGGDGDEKDKKKANLPKKEKSVLQGKLTKLAVQIGKAGLLMSAITVIILVLYFVIDTFWVQKRPWLAECTPIYIQYFVKFFIIGVTVLVVAVPEGLPLAVTISLAYSVKKMMKDNNLVRHLDACETMGNATAICSDKTGTLTMNRMTVVQAYINEKHYKKVPEPEAIPPNILSYLVTGISVNCAYTSKILPPEKEGGLPRHVGNKTECALLGLLLDLKRDYQDVRNEI
3	>6a69_B	Neuroplastin	Plasma membrane Ca2+-ATPase 1 (PMCA1) in	ALPHA-HELICAL	P-type ATPase	Homo sapiens	Human	MSGSSLPSALALSLLLVSGSLLPGPGAAQNEPRIVTSEEVIIRDSPVLPVTLQCNLTSSSHTLTYSYWTKNGVELSATRKNASNMEYRINKPRAEDSGEYHCVYHFVSAPKANATIEVKAAPDITGHKRSENKNEGQDATMYCKSVGYPHPDWIWRKKENGMPMDIVNTSGRFFIINKENYTELNIVNLQITEDPGEYECNATNAIGSASVVTVLRVRSHLAPLWPFLGILAEIIILVVIIVVYEKRKRPDEVPDDDEPAGPMKTNSTNNHKDKNLRQRNTN
4	>6a70_A	Polycystin-2	PKD2 polycystic kidney disease channel in lipid nanodiscs	ALPHA-HELICAL	Channels: Transient Receptor Potential (TRP)	Homo sapiens	Human	MGSAGWSHPQFEKGGGSGGGSGGSAWSHPQFEKGSAAAPRVAWAERLVRGLRGLWGTRLMEESSTNREKYLKSVLRELVTYLLFLIVLCILTYGMMSSNVYYYTRMMSQLFLDTPVSKTEKTNFKTLSSMEDFWKFTEGSLLDGLYWKMQPSNQTEADNRSFIFYENLLLGVPRIRQLRVRNGSCSIPQDLRDEIKECYDVYSVSSEDRAPFGPRNGTAWIYTSEKDLNGSSHWGIIATYSGAGYYLDLSRTREETAAQVASLKKNVWLDRGTRATFIDFSVYNANINLFCVVRLLVEFPATGGVIPSWQFQPLKLIRYVTTFDFFLAACEIIFCFFIFYYVVEEILEIRIHKLHYFRSFWNCLDVVIVVLSVVAIGINIYRTSNVEVLLQFLEDQNTFPNFEHLAYWQIQFNNIAAVTVFFVWIKLFKFINFNRTMSQLSTTMSRCAKDLFGFAI
5	>6a70_B	Polycystin-1	PKD2 polycystic kidney disease channel in lipid nanodiscs	ALPHA-HELICAL	Channels: Transient Receptor Potential (TRP)	Homo sapiens	Human	MGSAGDYKDHDGDYKDHDIDYKDDDDKGSAAATAFGASLFVPPSHVRFVFPEPTADVNYIVMLTCAVCLVTYMVMAAILHKLDQLDASRGRAIPFCGQRGRFKYEILVKTGWGRGSGTTAHVGIMLYGVDSRSGHRHLDGDRAFHRNSLDIFRIATPHSLGSVWKIRVWHDNKGLSPAWFLQHVIVRDLQTARSAFFLVNDWLSVETEANGGLVEKEVLAASDAALLRFRRLLVAELQRGFFDKHIWLSIWDRPPRSRFTRIQRATCCVLLICLFLGANAVWYGAVGDSAYSTGHVSRLSPLSVDTVAVGLVSSVVVYPVYLAILFLFRMSRSKVAGSPSPTPAGQQVLDIDSCLDSSVLDSSFLTFSGLHAEQAFVGQMKSDLFLDDSKSLVCWPSGEGTLSWPDLLSDPSIVGSNLRQLARGQAGHGLGPEEDGFSLASPYS
6	>1afo_A	GLYCOPHORIN A	Glycophorin A (GpA) transmembrane-	ALPHA-HELICAL	Glycoproteins	Homo sapiens	Human	VQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKK
7	>6agf_A	Sodium channel protein type 4 subunit alpha	NaV1.4-β1 complex	ALPHA-HELICAL	Channels: Potassium, Sodium, & Proton Ion-Selective	Homo sapiens	Human	MASWSHPQFEKGGGARGGSGGGSWSHPQFEKGFDYKDDDDKGTMARPSLCTLVPLGPECLRPFTRESLAAIEQRAVEEEARLQRNKQMEIEEPERKPRSDLEAGKNLPMIYGDPPPEVIGIPLEDLDPYYSNKKTFIVLNKGKAIFRFSATPALYLLSPFSVVRRGAIKVLIHALFSMFIMITILTNCVFMTMSDPPPWSKNVEYTFTGIYTFESLIKILARGFCVDDFTFLRDPWNWLDFSVIMMAYLTEFVDLGNISALRTFRVLRALKTITVIPGLKTIVGALIQSVKKLSDVMILTVFCLSVFALVGLQLFMGNLRQKCVRWPPPFNDTNTTWYSNDTWYGNDTWYGNEMWYGNDSWYANDTWNSHASWATNDTFDWDAYISDEGNFYFLEGSNDALLCGNSSDAGHCPEGYECIKTGRNPNYGYTSYDTFSWAFLALFRL
8	>6agf_B	Sodium channel subunit beta-1	NaV1.4-β1 complex	ALPHA-HELICAL	Channels: Potassium, Sodium, & Proton Ion-Selective	Homo sapiens	Human	MGRLLALVVGAALVSSACGGCVEVDSETEAVYGMTFKILCISCKRRSETNAETFTEWTFRQKGTEEFVKILRYENEVLQLEEDERFEGRVVWNGSRGTKDLQDLSIFITNVTYNHSGDYECHVYRLLFFENYEHNTSVVKKIHIEVVDKANRDMASIVSEIMMYVLIVVLTIWLVAEMIYCYKKIAAATETAAQENASEYLAITSESKENCTGVQVAE
9	>6ah4_A	P2X purinoceptor 3	ATP-gated P2X3 ion channel, closed apo state	ALPHA-HELICAL	Channels: Other Ion Channels	Homo sapiens	Human	SREFDFFTYETPKVIVVKSWTIGIINRVVQLLIISYFVGWVFLHEKAYQVRDTAIESSVVTKVKGSGLYANRVMDVSDYVTPPQGTSVFVIITKMIVTENQMQGFCPESEEKYRCVSDSQCGPERLPGGGILTGRCVNYSSVLRTCEIQGWCPTEVDTVETPIMMEAENFTIFIKNSIRFPLFNFEKGNLLPNLTARDMKTCRFHPDKDPFCPILRVGDVVKFAGQDFAKLARTGGVLGIKIGWVCDLDKAWDQCIPKYSFTRLDSVSEKSSVSPGYNFRFAKYYKMENGSEYRTLLKAFGIRFDVLVYGNAGKFNIIPTIISSVAAFTSVGVGTVLCDIILLNFLKGADQYKAKKFEEVNE
10	>4al0_A	PROSTAGLANDIN E SYNTHASE	Microsomal Prostaglandin E Synthase 1	ALPHA-HELICAL	Membrane-Associated Proteins in Eicosanoid and Glutathione Metabolism (MAPEG)	Homo sapiens	Human	MPAHSLVMSSPALPAFLLCSTLLVIKMYVVAIITGQVRLRKKAFANPEDALRHGGPQYCRSDPDVERCLRAHRNDMETIYPFLFLGFVYSFLGPNPFVAWMHFLVFLVGRVAHTVAYLGKLRAPIRSVTYTLAQLPCASMALQILWEAARHL
11	>4avm_A	BRIDGING INTEGRATOR 2	None	ALPHA-HELICAL	None	Homo sapiens	Human	SMGLFAKQVQKKFSRAQEKVLQKLGKAVETKDERFEQSASNFYQQQAEGHKLYKDLKNFLSAVKVMHESSKRVSETLQEIYSSEWDGHEELKAIVWNNDLLWEDYEEKLADQAVRTMEIYVAQFSEIKERIAKRGRKLVDYDSARHHLEAVQNAKKKDEAKTAKAEEEFNKAQTVFEDLNQELLEELPILYNSRIGCYVTIFQNISNLRDVFYREMSKLNHNLYEVMSKLERQHSNK
12	>4aw6_B	CAAX PRENYL PROTEASE 1 HOMOLOG	CAAX protease ZMPSTE24	ALPHA-HELICAL	Intramembrane Proteases	Homo sapiens	Human	MGMWASLDALWEMPAEKRIFGAVLLFSWTVYLWETFLAQRQRRIYKTTTHVPPELGQIMDSETFEKSRLYQLDKSTFSFWSGLYSETEGTLILLFGGIPYLWRLSGRFCGYAGFGPEYEITQSLVFLLLATLFSALAGLPWSLYNTFVIEEKHGFNQQTLGFFMKDAIKKFVVTQCILLPVSSLLLYIIKIGGDYFFIYAWLFTLVVSLVLVTIYADYIAPLFDKFTPLPEGKLKEEIEVMAKSIDFPLTKVYVVEGSKRSSHSNAYFYGFFKNKRIVLFDTLLEEYSVLNKDIQEDSGMEPRNEEEGNSEEIKAKVKNKKQGCKNEEVLAVLGHELGHWKLGHTVKNIIISQMNSFLCFFLFAVLIGRKELFAAFGFYDSQPTLIGLLIIFQFIFSPYNEVLSFCLTVLSRRFEFQADAFAKKLGKAKDLYSALIKLNKDNLGFPVSDWLFSMWHYSHPPL
13	>6awn_A	Sodium-dependent serotonin transporter	Serotonin transporter (SERT) ts3 construct with	ALPHA-HELICAL	Solute Carrier (SLC) Transporter Superfamily	Homo sapiens	Human	GSQGERETWGKKVDFLLSVIGYAVDLGNVWRFPYICAQNGGGAFLLPYTIMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIASYYNTIMAWALYYLISSFTDQLPWTSCKNSWNTGNCTNYFSEDNITWTLHSTSPAEEFYTRHVLQIHRSKGLQDLGGISWQLALCIMLIFTVIYFSIWKGVKTSGKVVWVTATFPYIALSVLLVRGATLPGAWRGVLFYLKPNWQKLLETGVWIDAAAQIFFSLGPGFGVLLAFASYNKFNNNCYQDALVTSVVNCMTSFVSGFVIFTVLGYMAEMRNEDVSEVAKDAGPSLLFITYAEAIANMPASTFFAIIFFLMLITLGLDSTFAGLEGVITAVLDEFPHVWAKRRERFVLAVVITCFFGSLVTLTFGGAYVVKLLEEYATGPAVLTVALIEAVAVSWFYGITQFCRDVKEMLG
			Serotonin					GSQGERETWGKKVDFLLSVIGYAVDLGNVWRFPYICAQNGGGAFLLPYTIMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIASYYNTIMAWALYYLIS

Figure 4. Alpha helical Transmembrane protein of Humans (Homo sapiens)

Methodology: The methodology involves the extraction of alpha-helical transmembrane proteins from the database PDBTM Then, feature extraction and selection were done and this information is transferred in a Microsoft EXCEL file which is then transferred to python software (Py-Charm). Python imports an excel file with the extension of CSV and also some directories to execute with processing such as the implementation of classifiers over dataset for designing the model. After processing, query protein was given to the model and obtains the desired outcome in form of a percentage. Figure 5 demonstrates the block diagram of the methodology .

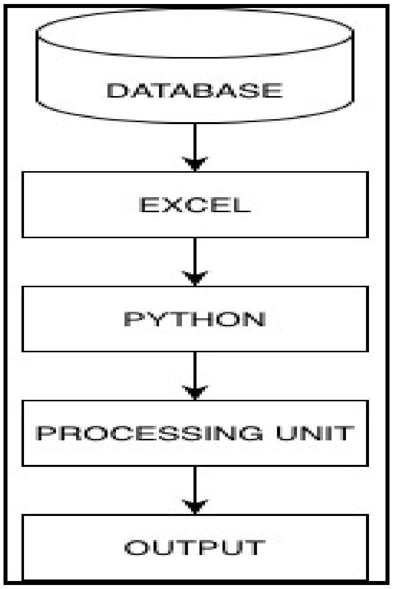


Figure 5. Block diagram of the methodology

Support Vector Machine: SVM is one of the widely used machine-learning algorithms for classification. It is a supervised machine learning approach that makes use of known inputs and outputs for training purposes. Once trained, now the algorithm can predict new inputs based on initial training and can differentiate between two groups. Before implementing SVM on human proteins, feature extraction is done. The extracted features are based on the amino acid composition of a protein and their sequential order as the number of amino acids varies in every transmembrane protein, we need to equalize the length of the protein. For this purpose, string balancing is done on the proteins.

Training and Testing: The database was divided into two chunks, one for training and the other for testing purposes. Training of the database was done by a support vector machine. After training, performance was evaluated by testing the dataset, the amino acid sequences were again given as input. Figure 6 depicts the flow diagram of the procedure.

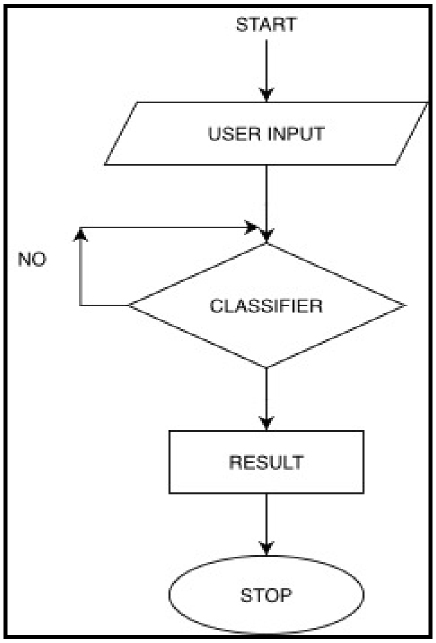


Figure 6. Flow diagram of the procedure

If the model is accurately trained, prediction in form of sequence similarity will be higher than 50% manifesting the presence of alpha-helical transmembrane human proteins.

RESULTS AND DISCUSSION

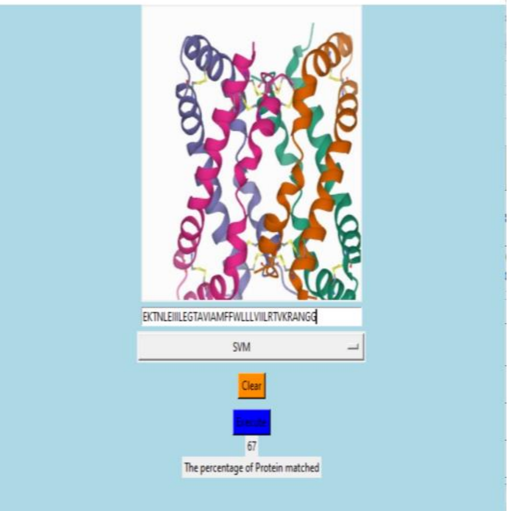
The results are displayed through Graphical User Interface (GUI), where protein sequence will be given as an input~~,~~ processed through the classifier and the percentage of alpha-helices is shown on the screen. Figure 7 and Figure 8 show the outputs for the alpha-helices in human proteins.

INPUT SEQUENCE

EKTNLEIIILVGTAVIAMFFWLLLVIILRTVKRANGG

A. Identity molecule =Vascular endothelial growth factor receptor 2
B. Protein name =VEGFR2 vascular endothelial growth factor receptor 2 transmembrane dimer
C. Organism = homo sapiens

OUTPUT



EKTNLEIIILEGTAVIAMFFWLLLVIILRTVKRANGG

SVM

67

The percentage of Protein matched

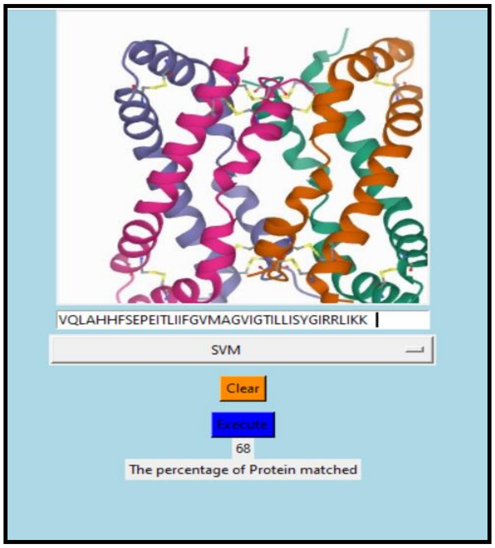
Figure 7. Output of first protein

Fig 7 reveals that the human protein with amino acid sequence EKTNLEIIILVGTAVIAMFFWLLLVIILRTVKRANGG contains 67% of alpha helices in its structural framework. It indicates that the given sequence is an alpha-helical protein found in human beings.

INPUT SEQUENCE:
VQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKK

A. Identity molecule = Glycophorin A
B. Protein name = Glycophorin A (GpA) transmembrane-domain dimer
C. Organism = homo sapiens

OUTPUT



VQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKK

SVM

68

The percentage of Protein matched

Figure 8. Output of second protein

Fig 8 depicts that the human protein with amino acid sequence VQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKK is an alpha-helical protein with 68% helices in its structural framework.

CONCLUSION

The main focus of this endeavor is to implement SVM uniquely in such a way that it can predict the amount of alpha-helices in human proteins. Unlike the traditional machine learning approaches, which only mention the accuracy of the predictor, we designed a graphical user interface (GUI) not only for those human proteins which are involved in training but also for new transmembane proteins.

For this purpose, we trained SVM by entering each and every protein sequence, so to memorize the similar pattern of helical sequences, which is then tested by entering the amino acid sequence in the graphical user interface (GUI). The percentage executed on the GUI depicts the folding level at the secondary level of the structure. More value in terms of percentage reveals the presence of more alpha-helices as compared to other forms i.e. beta sheets and hence, more spiral structure. Hence, a transmembrane protein with 88% on GUI shows the greater number of helical in its secondary structure as compare to that transmembrane protein for which output screen shows a value of 70%.

REFERENCES

Akram, M., Asif, M., Uzair, M., Naveed, A., Asadullah Madni, M., & Shah, M. A. Amino acids: A review article. *Journal of Medicinal Plant Research* , 3997–4000. (2011).

Apweiler, R. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, *29*(1), 37–40. (2001).

Barve, A., Ghaskadbi, S., & Ghaskadbi, S. Structural and sequence similarities of Hydra Xeroderma pigmentosum a protein to human homolog suggest early evolution and conservation. *BioMed Research International*, *2013*, 1–9. (2013).

Congreve , M., & Marshall, F. The impact of GPCR structures on pharmacology and structure-based drug design . *British Journal of Pharmacology*, *159*(5), 986–996. (2010).

Cai, C. Z., Yuan, Q. F., Xiao, H. G., Liu, X. H., Han, L. Y., & Chen, Y. Z. Prediction of Transmembrane Proteins from Their Primary Sequence by Support Vector Machine Approach . *Computational Intelligence and Bioinformatics*, 525–533. (2006).

Conn, P. M., Ulloa-Aguirre, A., Ito, J., & Janovick, J. A. G protein-coupled receptor trafficking in health and disease: Lessons learned to prepare for therapeutic mutant rescue in vivo. *Pharmacological Reviews*, *59*(3), 225–250. (2007).

Cong, H., Zhang, M., Zhang, Q., Gong, J., Cong, H., Xin, Q., & He, S. Analysis of structures and epitopes of surface antigen glycoproteins expressed in bradyzoites of*toxoplasma gondii*. *BioMed Research International*, *2013*, 1–9. (2013).

Dokholyan, N. V. Experimentally-driven protein structure modeling. *Journal of Proteomics*, *220*, 103777. (2020).

Das, B. B., Park, S. H., & Opella, S. J. Membrane protein structure from rotational diffusion. *Biochimica Et Biophysica Acta (BBA) - Biomembranes*, *1848*(1), 229–245. (2015).

Feng, S., Zhang, W., Yang, J., Yang , Y., & Shen, H. Topology Prediction Improvement of α-helical Transmembrane Proteins Through Helix-tail Modeling and Multiscale Deep Learning Fusion. *Journal of Molecular Biology*, *432*(4), 1279–1296 (2020).

Heijne, G. von. Principles of membrane protein assembly and structure. *Progress in Biophysics and Molecular Biology*, *66*(2), 113–139. (1996).

Hönigschmid, P., Breimann, S., Weigl , M., & Frishman, DAllesTM: predicting multiple structural features of transmembrane proteins. *BMC Bioinformatics.* (2020).

Koch, I., & Schäfer, T. Protein super-secondary structure and quaternary structure topology: Theoretical description and application. *Current Opinion in Structural Biology*, *50*, 134–143. (2018).

Kubota, T., Lacroix, J. J., Bezanilla, F., & Correa, A. M. Probing α-310 Transitions in a Voltage-Sensing S4 Helix. *Biophysical Journal*, *107*(5), 1117–1128. (2014).

Landreh, M., & Robinson, C. V. A new window into the molecular physiology of membrane proteins. *The Journal of Physiology* , 355–362. (2014).

Murray, J. E., Laurieri, N., & Delgoda, R. Proteins. *Pharmacognosy*, 477–494. (2017)

Rask-Andersen, M., Almén, M. S., & Schiöth, H. B. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery*, *10*(8), 579–590. (2011).

Singh, R., Jain , N., & Pal Kaur, D. GOR Method for Protein Structure Prediction using Cluster Analysis . *International Journal of Computer Applications*, *73*(1), 1–6. (2013).

Shukla, H. D., Vaitiekunas, P., & Cotter, R. J. Advances in membrane proteomics and cancer biomarker discovery: Current status and future perspective. *PROTEOMICS*, *12*, 3085–3104. (2012)

Sun, J., & Frishman, D. Improved sequence-based prediction of interaction sites in α-helical transmembrane proteins by deep learning. *Computational and Structural Biotechnology Journal*, *19*, 1512–1530. (2021).

Tsai, C.-J., Maizel, J. V., & Nussinov, R. Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proceedings of the National Academy of Sciences*, **97**(22), 12038–12043. (2000).

Tenreiro Machado, J. A., Costa, A. C., & Quelhas, M. D. Can power laws help us understand gene and proteome information? *Advances in Mathematical Physics*, *2013*, 1–10. (2013).

Wang, L., Zhong, H., Xue, Z., & Wang, Y. Improving the topology prediction of α-helical transmembrane proteins with deep transfer learning. *Computational and Structural Biotechnology Journal*, *20*, 1993–2000. (2022)

Publisher's note: PJBT remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.