SSBI 2019: Report for Project 3

Luca Deininger, Nantia Leonidou, Alexander Ott

Transmembrane proteins play a vital role in living cells by allowing communication and transport between the cell and the outside environment. Due to their role in a vast number of physiological processes, membrane proteins are the most popular class of drug targets. Therefore distinguishing whether a protein is transmembrane or not is an important scientific question. The following report presents a classifier which, written in Python 3.7.2., determines whether a given protein structure is of an alpha-helical transmembrane protein, and if so, can identify the approximate membrane plane and the transmembrane helices.

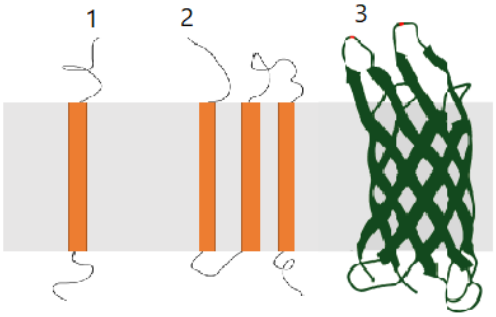

Fig. 1. Schematic representation of transmembrane proteins. Helices are coloured with orange, while green represents beta barrels.

Introduction

Biological membranes are essential for cells and are highly involved in cellular activities. They allow separation between the inside and outside of a cell and act as selectively permeable barriers. This enables the cell to simultaneously take up essential substances (e.g. gases, solutes and more) from the external environment and eliminate toxic products from the interior (1). Their composition was first described by Singer and Nicolson in 1972 as the so-called fuid-mosaic model (2). According to this model, they consist of a phospholipid bi-layer with embedded, integral and peripheral proteins. Membrane proteins play a very important role, as they maintain the organization and flow of material through the membranes. The associated amino acids are positioned based on their polarity. Non-polar amino acids are located on the inside and are in direct contact with water solutions, while non-polar residues are facing the lipid bi-layer. Different types of membrane proteins incorporated into the lipid bi-layer provide different characteristics of the bio-membranes. Membrane proteins can be distinguished in two main categories: peripheral or integral. Peripheral membrane proteins, also called water-soluble globular proteins, are temporarily fixed to one side of the membrane via non-covalent interactions. They can be bound either to phospholipids or to the surface of integral proteins. On the other hand, integral membrane proteins (IMPs) are permanently anchored to the membrane and penetrate its hydrophobic interior. They are fixed stronger to it and are harder to remove, only by destroying the membrane. They form about $20-30\%$ of all protein sequences and are very important for transport and communication between cells and their external environment (3). IMPs can be classified to integral monotopic and polytopic proteins. Monoptotic proteins do not span the entire lipid bi-layer, while polyptotic (transmembrane proteins) are extended through the whole membrane.

Transmembrane proteins. As mentioned above, transmembrane proteins (TMPs) span the entire membrane and parts of them are exposed on both internal and external environment. They are usually larger than non-transmembrane (globular) proteins, making the determination of their structure and crystallization quite difficult (3). There are two kinds of TMPs: alpha-helical and beta-barrels. TMPs can be either single transmembrane alpha helices (bitopic), as you can see in Figure 1 first tertiary structure, or they can be composed of many alpha helices bundled together (Fig. 1, No2). This type of TMPs can be usually found inside bacterial cells or plasma membrane of eukaryotes. Beta barrels (Fig. 1, No3) are mostly common in channel proteins and can be found in mitochondria and chloroplasts. The 3D structure of TMPs can be obtained by NMR spectroscopy or X-ray crystallography and is regarded as one of the most difficult tasks, compared to peripheral proteins. Determining their structures involves their removal from lipid bi-layer, thus the membrane is missing from the generated structures and structural information about the proteins is hard to obtain. The topology of TMPs can be easily predicted from the

sequence of amino acids. The protein part, that is attached to the lipid bi-layer, has a high consistency of hydrophobic amino acids. However, this kind of prediction is not adequate to distinguish between TMPs and globular proteins. Using as input their 3D structure would deliver more information and would enable the identification of TMPs and their location (3, 4).

The Protein Data Bank (PDB) contains only soluble proteins and does not differentiates between TMPs and non-TMPs (5). The file description might implies that some proteins are TMPs, but those kind of information is very hard to extract direclty. Thus, the goal of the implemented tool is to determine whether a given PDB structure is an aplha-helical TMP, and if so, to identify the approximate membrane plane and the transmembrane helices. The model was validated using structures from the Protein Data Bank of Transmembrane Proteins as reference standard (PDBTM) (6).

Materials and Methods

The data and methods used to investigate differences of transmembrane (TM) and non transmembrane (non-TM) helices, for the construction of a reasonable classifier and the validation of this classifier is described in the following.

Investigation of differences in amino acid composition of TM and non-TM helices. In order to build a classifier for TM helices, differences in amino acid composition of TM and non-TM helices were investigated.

Data. Helices of a PDB structure were extracted with DSSP, like explained in the next section. Sequences of 1332 non-TM helices from 100 randomly sampled PDB structures, not contained in the PDBTM database were extracted. In order to derive sequences of TM-helices, the xml file of all PDBTM entries (6) was parsed. From the complete xml file, 57.345 TM alpha-helices were extracted. TM and non-TM helices were analyzed concerning their amino acid composition.

Obtaining secondary structures of PDB proteins. Dictionary of Secondary Structure of Proteins (DSSP) is an algorithm that assigns secondary structures to amino acids of a protein based solely on backbone-backbone hydrogen bonds (7). Whether an H-bond exists is determined using the electrostatic energy for each acceptor/donor pair given by the formula:

$$E = q_1 \cdot q_2 (\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}) \cdot f$$

(7) where $q_1 = 0.42e$, $q_2 = 0.20e$, e is the unit electron charge and r_{ij} are the distances between atoms i and j (in Angstroms). The factor f is equal $332 \frac{\text{kcal}}{\text{mol}}$. The process can assign eight different types of secondary structures: 3_{10} helices (**G**), helix (**H**), π-helix (**I**), beta bridge (**B**), beta bulges (**E**), turns (**T**), bends (**S**) and if no rule applies a blank space (**-**) is written. The DSSP() function handles one model and DSSP data is accessed by a tuple of chain and residue IDs. For the purpose of extracting helices of a given PDB structure, the DSSP output is parsed by extracting amino acids with assigned secondary structure H which are subsequently chained into separate helices.

Classification of helices based on physico-chemical properties of helices. We used a Support vector machine (SVM) as a classifier to distinguish between TM and non-TM helices. Support vector machines are a type of supervised machine learning classification algorithm. Its main goal is to find a hyperplane that maximizes the margin between two object classes (8). This was done using the `svc()` library from the scikit-learn package (9). Some of the advantages of SVMs are their effectiveness in high dimensional spaces and that many different kernel functions can be used as decision functions (9). Since a classification needs to be performed, a support vector classifier (SVC) was used which is an implemented function under the name `svc()` in Scikit-learn module. The actual training and the parametrization of the SVMs is explained in the two following subsections.

Training. Since the investigation of differences in amino acid composition of TM and non-TM helices showed significant differences between both classes, SVMs were trained on counts of each proteinogenic amino acid. For the training of SVMs, sequences of all 1332 non-TM helices and 3000 randomly sampled sequences of TM-helices, obtained from the previous task of investigating differences in acid composition, were used. Only 3000 out of 57.345 TM alpha-helices were used to keep the ratio of TM and non-TM helices in a reasonable range, but still high enough to

learn specific features of TM helices. For this purpose, each helix of the previously derived data was labeled with 1 if the helix is TM, and 0 if the helix is non-TM. For training, each helix sequence was converted into an array containing the counts of each proteinogenic amino acid in the corresponding helix sequence. Three SVMs were trained in a different manner. *SVM_abs* was trained on absolute counts of each proteinogenic amino acid, while *SVM_rel* was trained on relative frequencies of each proteinogenic amino acid. Additionally, the xml file of all PDBTM entries (6) was parsed to extract sequences of TM non alpha-helix secondary structures. 3000 of their sequences were labeled as non-TM and included in the dataset for training of *SVM_abs_ext*. This was done to provide the SVM with further features that might allow the SVM to better learn TM alpha-helix specific features. The absolute counts of each proteinogenic amino acid were used for training of *SVM_abs_ext*.

Parametrization. A linear kernel with a penalty parameter C equal to 1.0 (10) was used, since adaptions of C in a range of 1.0 to 5.0 lead to negligible differences in TPR and FPR of smaller than 0.1%. The parameter C informs about how much we want to avoid missclassification. A large C value would lead to small margin, while small C values lead to lower missclassification rate. For actual training the fit_proba() method was called to train the SVM on the training data. fit_proba() was used instead of fit() to derive confidence scores for each prediction of the SVM for later classification.

Refinement of the classification. To improve the annotation coming from the SVM, the annotations were analyzed, filtered and the membrane position and orientation was estimated to further refine the result and increase the precision.

Analysis and filtering. To increase the quality of the annotations used for the estimation of the membrane, proteins that had only one annotated TM-helix were treated as globular and proteins with less than 10 TM-helices need an average confidence of the classifier higher than 90% to be counted as true hits. If the conditions were not hit, the annotation as TM was removed and protein classified as globular.

Estimation of membrane position and orientation. For proteins that fulfilled the conditions, the membrane was approximated. This work is split into two steps: the approximation of the membrane normal and the approximation of the position. To calculate the normal of the membrane first the axis of all TM helices was calculated using the method described in Kahn, 1989 (11). Using adjacent pairs of Cα's in the helix backbone the axis direction of this part is defined. The axis of the whole helix is then calculated using singular value decomposition (SVD) on the directions of all parts (see below) and the normal of the membrane was then approximated using SVD on the directions of all TM helices in the protein. The position of the membrane was approximated using the average position of the middle Cα of the helices. Singular value decomposition (SVD) is a method of decomposing a matrix into three other matrices (two orthogonal and one diagonal) (12). From this, one can compute the singular values of the matrix. These values characterize, similar to the eigenvalues, properties of the matrix. Singular value decomposition exist for every matrix (square or non-square matrices). Then the singular value decomposition of a matrix M is the factorization of M into the product of three matrices: $M = UDV^T$, where M is an $m \times n$ matrix, U is an $m \times n$ orthogonal matrix, D an $n \times n$ diagonal matrix and V an $n \times n$ orthogonal matrix (13). The diagonal entries in D are defined as the singular values of M.

Refinement of annotation with approximated membrane. After the estimation of the membrane position and orientation, a final refinement step is taking place, to filter out helices that do not intersect the membrane and thus do not involve in a membrane plane. For each helix it is tested if it has Cα atoms on both sides of the membrane. Depending on whether the helix is intersecting with the membrane the annotation is changed.

Annotation of helices and validation.

Data. In order to evaluate the performance of our classifier and subsequent refinement, 500 PDB structures, not contained in the PDBTM, and 100 PDB structures, contained in the PDBTM, were sampled randomly. Of all PDB structures, 13.897 helices were extracted with DSSP, in a manner like previously explained.

Annotation with PDBTM. All extracted helices were annotated with the PDBTM representing the correct

annotation of each helix of all PDB structures. If a PDB structure was not contained in the PDBTM, all its helices were annotated as non-TM. If a PDB structure was contained in the PDBTM, every helix was looked up in the PDBTM xml file whether the corresponding helix is annotated as TM. For this purpose, the chain and start and end residue identifiers of both helices were compared. Observing that sometimes the start and end residue identifiers of a TM helix in the PDBTM and the definition of the corresponding helix obtained from DSSP is different, it was necessary to define a minimal overlap of the PDB helix and the PDBTM helix above a PDB is annotated as TM. We decided that at least half of the PDB helix needs to be covered by the PDBTM helix so that a PDB helix is annotated as TM. This indicates, that at least 50% of each TM helix lies inside the membrane. The definition of a general overlap fraction involves the risk of annotating helices wrongly, but this is unavoidable, since the start and end of helices of the PDBTM are not consistent with start and end of helices in the PDBTM. Analogously, the overlap of the PDBTM helix by the PDB helix was defined. This was necessary, since a TM helix that has no overlap with the PDB helix, annotated from DSSP, is in general unlikely a helix and thus should not be considered as a TM helix. Overall the annotation lead to 1.985 TM helices and 11.912 non-TM helices.

Annotation with SVM classifier. Each trained SVM was used to annotate helices and classify them to TM and non-TM helices. Given a trained SVM, each helix gets predicted by calling the function *predict_proba()* from the scikit-learn package (9).

Refinement with Membrane. As discussed above the refinement was used to improve the precision of the annotations.

Validation. We compared our annotation with the PDBTM annotation and calculated the true positive rate (TPR), false positive rate (FPR), the ratio of correctly classified helices divided by the total number of helices and the total number of correctly classified proteins after SVM classification and after subsequent refinement. A protein is classified as TM if it contains at least one TM helix. Additionally, we performed a 10-fold cross-validation with the scikit-learn package (9) on the data set, used for training of the SVMs.

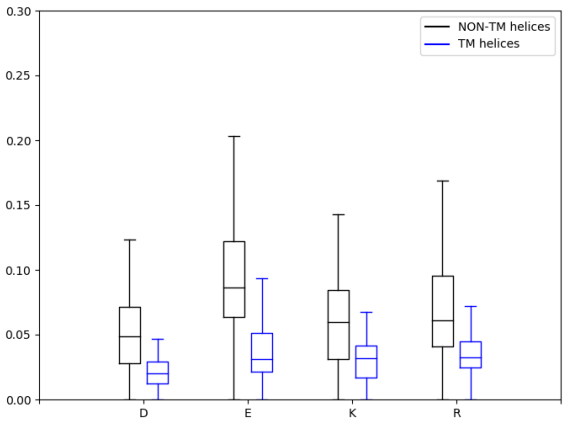

Fig. 2. Distributions of frequencies of hydrophilic amino acids D, E, K, R in TM vs non-TM helices.

Results and Discussion

In the following, the results concerning the investigation of differences in amino acid composition of TM and non-TM helices, the validation of our built classifier and the membrane approximation are presented and discussed.

Differences in amino acid composition of TM and non-TM helices. TM helices are considered to contain higher proportions of hydrophobic and lower proportions of hydrophilic amino acids, thus the frequency of amino acids in TM and non-TM helices was investigated in more detail. 1332 non-TM helices from 100 randomly sampled PDB structures, were compared to 3000 TM helices extracted from the PDBTM. The distribution of frequencies for hydrophobic and hydrophilic amino acids are shown in Fig. 2 and 3. It can observed that hydrophilic amino acids acids D, E, K and R are significantly underrepresented in TM helices, as you can see in Figure 2, while hydrophobic amino acids, especially F and G, are over-represented in TM helices (see Fig. 3).

Validation of classifier. Since the amino acid composition of TM and non-TM helices showed significant differences, the idea was to train SVM classifiers on amino acid frequencies. For validation, the annotation of this classifier further combined with the refinement was compared to the annotation of PDBTM (for details see Materials and Methods). Table 2 before refinement shows

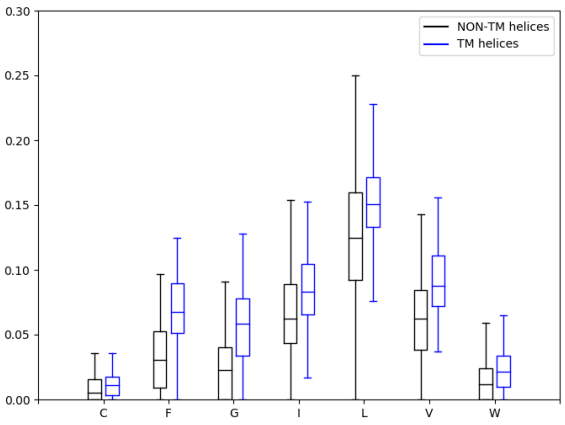

Fig. 3. Distributions of frequencies of hydrophobic amino acids C, F, G, I, L, V and W in TM vs non-TM helices.

that SVM_abs and SVM_abs_ext perform similarily well. While SVM_abs shows higher TPR, SVM_abs_ext reached lower FPR. However, the higher number of correctly classified helices, suggests a better performance of SVM_abs_ext. Indicating that the usage of sequences of TM non-helix secondary structures labeled as non-TM for training of SVM_abs_ext allowed a better classification. Surprisingly, SVM_rel shows significantly higher FPR and lower classification rate compared to both other SVMs, however its performance enhanced a lot after refinement. It stands out, that after refinement TPR decreased for all SVM classifiers while the FPR decreased as well. Furthermore, the number of correctly classified helices increased for all SVMs, especially for SVM_rel. Overall our classifier SVM_abs_ext with subsequent refinement performs best with over 95% correctly classified helices. Comparing Table 1 and 2, it strikes that the cross-validation of all SVMs showed better performance for the training data set compared to the test data set. That can be explained by inconsistencies between helices defined in the PDB and PDBTM (for details see "Annotation with PDBTM", Materials and Methods) leading to an increase of wrongly annotated helices by PDBTM and thus a little decrease in the number of correctly classified helices for the test data set.

Case study of incorrectly classified TM proteins. From the 21 not correctly classified proteins (see Table 2 SVM_abs_ext), 12 were false negatives and nine false positives. A further investigation showed

Table 1. Mean and standard deviation of correctly classified helices for 10-fold cross-validation for all SVMs on training data set.

	SVM_abs	SVM_abs_ext	SVM_rel
μ (%)	96.14	94.59	93.9
σ (%)	1.15	1.41	1.53

Table 2. TPR, FPR and correctly classified helices of validation data set for all SVMs after SVM classification (top) and after additional refinement (bottom). For bottom table the number of correctly classified proteins is indicated.

After SVM classification	SVM_abs	SVM_abs_ext	SVM_rel
TPR (%)	95.31	90.98	93.09
FPR (%)	9.57	6.30	19.93
Correctly classified helices (%)	91.12	93.30	82.03
After refinement			
TPR (%)	81.46	81.86	76.32
FPR (%)	3.23	2.65	5.56
Correctly classified helices (%)	94.58	95.13	91.84
Correctly classified structures	557/583	562/583	505/583

that nine of 12 false negative proteins were TM proteins only containing beta-sheets. As our classifier is designed to detect TM alpha-helices, this is not considered as a weakness of our classifier. In the following was taken a closer look on the three incorrectly classified alpha-helical TM proteins (PDB-IDs: 2j58, 3t2m and 6djr). For 2j58, zero out of eight TM helices were classified as TM. A further investigation of the amino acid composition of these helices (all helices were identical) showed surprisingly high abundances of serine and threonine (each 3 out of 17 amino acids). Considering the high confidence scores of the SVM_abs prediction of 81.99% for each helix, the unusual amino acid composition of these helices is the most probable cause of this misclassification. For the remaining two structures, trivial reasons could be found. The structure 3t2m is different in the PDB and PDBTM, consequently leading to a wrong annotation of the structure by PDBTM. 6djr is a TM structure consisting of arbitrary amino acids X only, allowing no amino acid based classification.

Validation of membrane approximation. To validate the approximated values for the normal and the position of the membrane, we used the xml files provided on PDBTM. The position was validated using the distance of our position to the plane of defined in the xml (see Fig. 4). The distance was surprisingly small for such a simple approach with most distances between 0 and 0.5 $\mathring{A}$ and only few

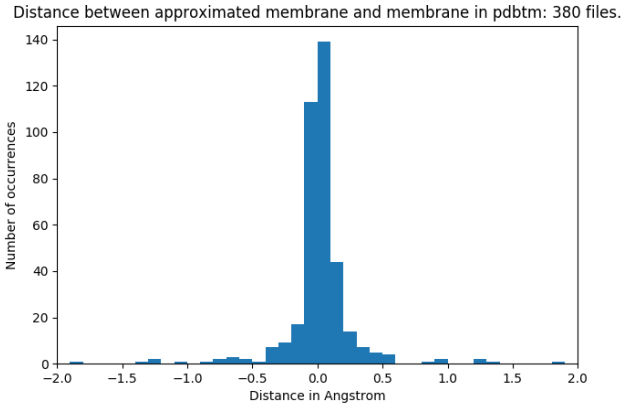

Fig. 4. The distance between the approximated position of the membrane and the membrane plane defined in the PDBTM.

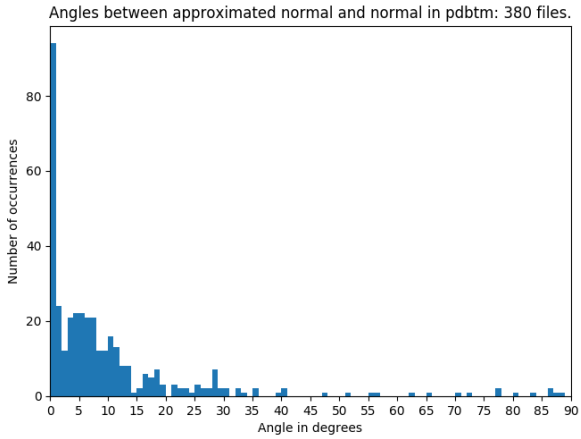

Fig. 5. The Angle between the approximated membrane normal and the normal given in PDBTM. Most approximations fit to the PDBTM normal within 10 degrees but there are outliers up to 90 degrees.

outliers. The approximation of the normal of the membrane on the other hand has a lot more outliers but most values have less than a 15 degree angle to the reference angle from the PDBTM (see Fig. 5). This was in part caused by our approach to the refinement which increased the precision for the annotation of the helices but introduced a lot more outliers in the approximation of the normal.

To steer against this increase it may be possible to introduce additional constraints into the refinement process that reduce the possibility to erroneously refine away from the true membrane. One possibility of such a constraint may be that the annotation with very high confidences from the classifier can't be changed.

Conclusion and Outlook. The validation of our classifier shows it performs well, with over 95% correctly identified helices and proteins. Inconsistencies of helix annotations of the PDB and PDBTM are a source of error leading to misclassifications as well as TM helices with unusual amino acid composition. In order to build an even more powerful classifier, further features of TM helices beyond specific amino acid composition could be considered as well. The paper of Hildebrand et al. presents further specific properties of TM helices, e.g. differences in mean of phi and psi angles compared to non-TM helices (14). Nevertheless the here presented classifier uses the information of amino acid frequencies only and shows this simple approach performs, together with a refinement step using the approximated membrane, reasonably well.

References.

1. Stillwell W (2016) *An introduction to biological membranes: composition, structure and function.* (Elsevier).
2. Singer SJ, Nicolson GL (1972) The fluid mosaic model of the structure of cell membranes. *Science* 175(4023):720–731.
3. Tusnády GE, Dosztányi Z, Simon I (2004) Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics* 20(17):2964–2972.
4. White SH, Wimley WC (1999) Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure* 28(1):319–365.
5. Berman HM, et al. (2000) The protein data bank. *Nucleic acids research* 28(1):235–242.
6. Kozma D, Simon I, Tusnady GE (2012) Pdbtm: Protein data bank of transmembrane proteins after 8 years. *Nucleic acids research* 41(D1):D524–D529.
7. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.
8. Bishop CM (2006) *Pattern recognition and machine learning.* (springer).
9. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.
10. Buitinck L, et al. (2013) API design for machine learning software: experiences from the scikit-learn project in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning.* pp. 108–122.
11. Kahn PC (1989) Defining the axis of a helix. *Computers Chemistry* 13(3):185 – 189.
12. Klema V, Laub A (1980) The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control* 25(2):164–176.
13. Gentle J (1998) Singular value factorization. *Numerical Linear Algebra for Applications in Statistics* pp. 102–103.
14. Hildebrand PW, Preissner R, Frommel C (2004) Structural features of transmembrane helices. *FEBS Lett.* 559(1-3):145–151.