Identifying Transmembrane Helices in Tertiary Structures

Marco Klein, Julian Späth, Julius Vetter

marco.klein@student.uni-tuebingen.de
julian.spaeth@student.uni-tuebingen.de
julius.vetter@student.uni-tuebingen.de
University of Tuebingen

I. Introduction

Transmembrane (TM) proteins play an important role in many physiological processes in the cell, such as the transport of molecules and ions or cell signalling. Most commonly, TM proteins contain either a helix bundle or beta barrel substructure which is embedded in the membrane [1].

However, the ratio of experimentally determined tertiary structures of TM proteins is rather low. Furthermore, even if the structure of a TM protein is determined, an important detail, its topology relative to the lipid bi-layer, is not annotated in the PDB. Therefore, the computational prediction of TM substructures such as helix bundles or beta barrels, and the subsequent derivation of the membrane geometry for TM proteins is a relevant biological question.

Various methods have been proposed for this task. Rost *et al.* [2] apply a neural network approach, integrating evolutionary information from related sequences into the helix classification. A hidden Markov model is utilised by Krogh *et al.* [3] to model the topology of TM proteins with 97% accuracy. Both methods use sequence information as their input. Tusnády *et al.* [4] propose the TMDET algorithm to score the viability of a membrane geometry for a protein, which divides the tertiary structure into slices along the tested orientation and calculates hydrophobicity and structure factors for each slice.

The focus of this project lies on predicting the transmembrane helices of α-helical TM proteins to help determine the geometry of a membrane. For this purpose, a support vector machine (SVM) has been implemented which predicts whether a helix is transmembrane or soluble from its amino acid composition. Furthermore, a method has been implemented which, given the predicted TM helices of a protein, calculates the tentative position and orientation of the membrane. This initial position and orientation is then refined using the objective function of the TMDET algorithm.

II. Methods

Data Processing

Procuring a dataset of high quality is essential for the success of any classifier. To distinguish between soluble and TM helices, information has to be extracted from annotated transmembrane regions. This data was retrieved from the Protein Data Bank of Transmembrane Proteins [4] (PDBTM), which lists over 3,000 known TM proteins and their respective topologies (marking intra- and extracellular, as well as transmembrane regions).

In order to extract TM helices, transmembrane regions of α-helical PDBTM entries have been matched with α-helix regions of their respective files from the Protein Data Base [5] (PDB). As the start and end positions of TM regions annotated in the PDBTM do not perfectly match those of their respective helix annotations of the PDB, e.g. because only a segment of the TM helix is embedded in the membrane, an overlap has been introduced to determine

TM helices. If a TM segment of a PDBTM entry and a helix segment of the corresponding PDB entry have an overlap of at least 70%, this helix will be considered a TM helix, i.e. a positive sample for the predictor. Helices of PDBTM proteins with 0% overlap were labelled as negative samples. Finally, 14,356 α-helices from globular proteins were added to the dataset as additional non-TM samples.

The information extracted from each sample helix was its position (chain, start and end residues) and the amino acid sequence of the helix. In the end, the dataset contained 83,388 fully annotated helices of which 35,059 were transmembrane.

Helix Prediction

The implementation of the transmembrane helix predictor consisted of two steps: the encoding of the features of the dataset presented in section II and the training of a machine learning model. As both amino acid frequencies (Figure 1) and helix lengths (Figure 2) differ significantly between transmembrane helices and non-transmembrane helices, the features used to train the classifier were vectors of length 20 containing the frequency of each amino acid in a helix. Since absolute frequencies were used, this implicitly also encodes the helix length as a feature.

Using these features, support vector machines (SVM) with linear and rbf kernels have been trained as prediction models that classify helices as transmembrane and non-transmembrane in a binary manner. The predictor was implemented in Python 3.6 using the scikit-learn [6] framework (0.19.1).

The dataset has been split before training into a training set (70%) and test set (30%). This enabled a precise evaluation of the trained model on the independent the test set which had never been seen by the model before.

In order to remove the influence of the differently sized sample classes, the SVMs used a balanced class weight during training. The regularisation parameter C and the rbf kernel specific parameter γ were optimised by grid search using 5-fold cross validation and aiming for the highest Matthews correlation coefficient (MCC), a performance measure superior to plain accuracy on unbalanced datasets. The grid search evaluation resulted in an SVM with $C = 100$ and an rbf kernel with $\gamma = 0.1$.

Membrane geometry calculation

The results of the helix classification for a certain protein were then used to estimate the position and orientation of the lipid membrane for transmembrane proteins. In order to determine the viability of a certain membrane orientation for a TM protein, an objective function called the Q-value as proposed by Tusnady *et al.* [4] was implemented in Python.

For a given set of helices classified as transmembrane, a two-step filtering was applied in order to remove false positive helices that could potentially corrupt the final membrane orientation. Firstly, a principal component analysis (PCA) was applied to the direction vectors of all helices, and helices that deviated more than 20° from the first PCA loading vector were removed.

As multiple TM helices cluster in the same region of the membrane, the second filtering step considered the position of the helices. To achieve this, all remaining helix centres were projected onto the first loading vector, and the resulting values were hierarchically clustered using the single linkage method. The clustering was terminated once a distance of 20 Å had been reached, and the cluster containing the largest number of sequences was deemed the "membrane helix cluster". All other predicted helices outside of this cluster were discarded. In the end, the orientation of the membrane is defined be the first PCA loading vector of the remaining helices which serves as the normal vector of the membrane. Its position is defined as the mean position of the remaining helix centres.

This tentative geometry is then optimised using the Q-value objective function. The Q-value is calculated for 1 Å slices of the protein tertiary structure, orthogonal to the tested

orientation vector. It is the product of a hydrophobicity and a structure factor. The hydrophobicity factor is defined as the ratio of the hydrophobic solvent accessible surface area (SASA) of all outer atoms, i.e. atoms that would interact with a potential membrane within a given slice. The structure factor rewards the "straight" residues in a slice and penalises "turn" residues as well as chain end residues. A residue i is defined as straight if the projection of the residues $(i-3)$, i, and $(i+3)$ onto the tested vector are monotonically increasing or decreasing. Otherwise, it is classified as a "turn" residue [4].

The algorithm only uses the previously calculated normal vector of the tentative membrane as an input. 30 uniformly distributed unit vectors are sampled in a 15° cone around the normal vector, and for each vector the Q-value distribution along its direction is computed. The vector which yields the largest Q-value average for a sliding window of 30 Å is chosen as the final normal vector for the membrane. The membrane position is defined by the location of the optimal window. All helix vectors before and after each filtering step, as well as calculated membrane geometries before and after optimisation, are exported as a JSON file. An additional script was created to use such a result file to visualise the protein-membrane topology in PyMol.

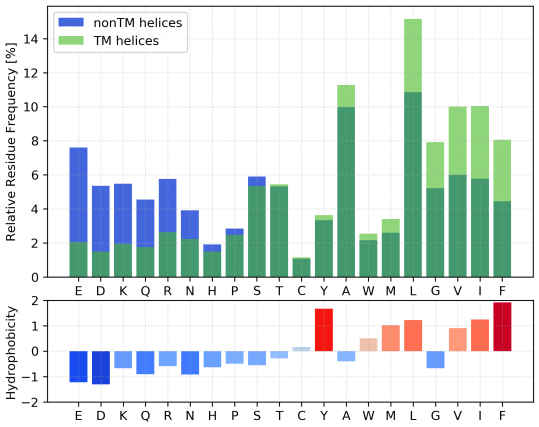

Figure 1: *Comparison of amino acid frequencies between transmembrane helices and non-transmembrane helices. Hydrophobicity scores from [7].*

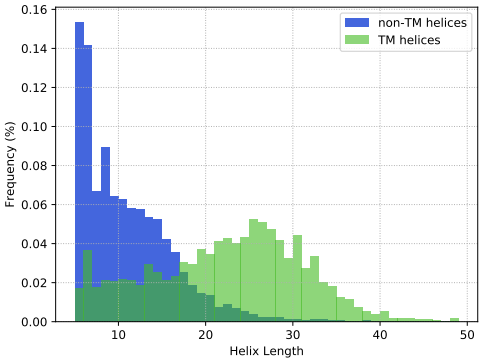

Figure 2: *Helix length distributions of transmembrane helices and non-transmembrane helices.*

III. Results

Analysing the dataset already gave insights into the differences of transmembrane and non-transmembrane helices. Figure 1 shows the differences of the amino acid frequencies of both classes. It can be seen that some amino acids preferentially occur in TM helices while others rather occur in non-TM helices.

Additionally, there are significant differences in the helix lengths of the two classes. As shown in Figure 2 transmembrane helices mostly contain between 15 and 30 amino acids and therefore are relatively large. On the contrary, non-transmembrane helices rarely consist of more than 17 amino acids.

Considering these exploratory insights, using amino acid frequency features for the classification seemed to be sufficient to achieve a highly accurate model. The trained SVM with optimal parameters was evaluated on an independent test set (30% of the original data) containing 10,421 positive and 14,596 negative samples. The confusion matrix of the prediction on the test data is shown in Table 1. It shows that 96% of transmembrane helices are correctly classified, and so are 96% of the negative samples. This corresponds to an f1-score of 0.95 and an MCC of 0.92. As an MCC of 0 represents random classification and 1 a perfect prediction, the score of 0.92 is very high and the predictor is close to perfect.

Table 1: *Confusion Matrix summarising the performance of the transmembrane helix predictor.*

	Actual TM	Actual non-TM
Pred. TM	TP = 10016	FP = 569
Pred. non-TM	FN = 405	TN = 14027

Membrane calculation

A manual validation of the filtering of false positive helices showed that falsely classified helices could be successfully removed from the TM helix set before geometry computation. There were some cases where the filtering removed correctly classified helices for the membrane calculation. However, this did not influence the resulting membrane geometries in any of the tested examples due to the subsequent optimisation step.

The influence of the optimisation procedure is visualised and interpreted in Figure 3 with the bovine cytochrome BC1 complex (1BE3). While the non-optimised membrane orientation (B) seems to be reasonable, there is one major difference to the optimised geometry which can be explained by the nature of the objective function. The partial helix marked in black is parallel to the membrane, and therefore more likely to contain "turn" residues which are penalised under the Q-score. Furthermore, this segment contains more hydrophilic than hydrophobic residues, leading to a lower hydrophobicity factor. In conclusion, the new geometry seems to be better suited for a membrane model.

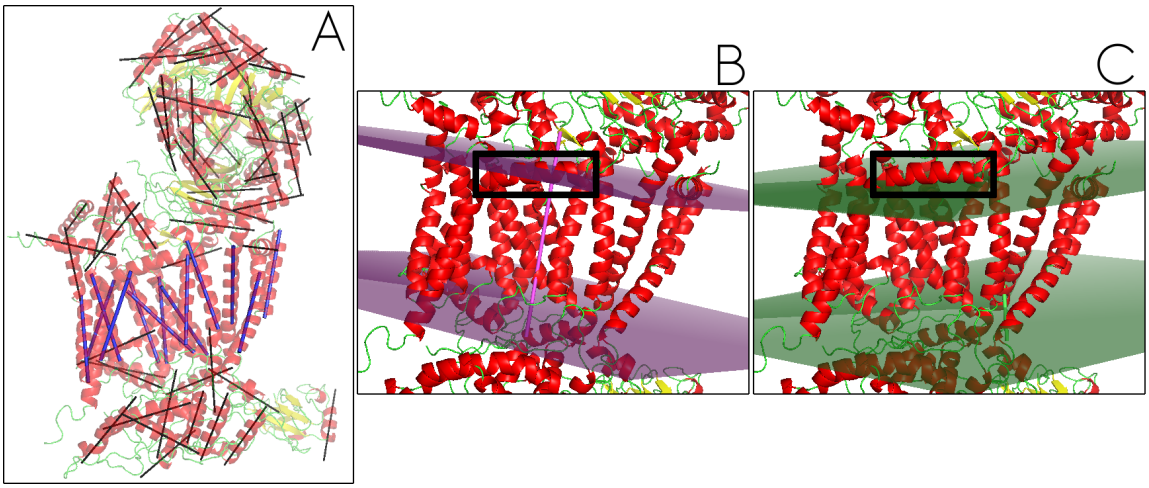

Figure 3: *Results of the helix prediction and membrane calculation of 1BE3. A: Helix prediction results. Blue lines represent helices predicted as TM, black lines represent non-TM predictions. B: Membrane geometry before optimisation. C: Membrane geometry after optimisation. The black border marks the helix discussed in this section.*

IV. Discussion

Despite having a helix and membrane prediction with such high accuracy, it is worth discussing how the SVM classifies and in what cases it would yield a sub-optimal membrane.

To get a better understanding on how the SVM classifies the helices, an SVM with linear kernel was trained on the same dataset. While its accuracy was slightly worse than the SVM with the rbf kernel (accuracy: 0.91, f1: 0.89, MCC: 0.81), it allows a direct interpretation of the trained weight vector.

As seen in Figure 4, the entries in the obtained weight vector closely resemble the hydrophobicity scores of the respective amino acids (Pearson $r = 0.87$). Hence, the linear SVM has learned to primarily consider the hy-

drophobicity of the helix for classification, despite never having had access to explicit hydrophobicity figures.

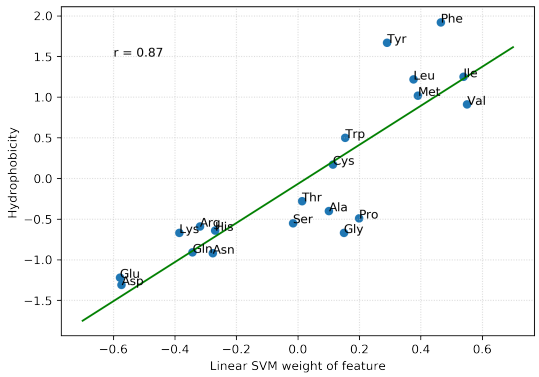

Figure 4: *Relationship between the weights of the linear SVM and the hydrophobicity scores of its features. The green line visualises the linear regression.*

While the classifier performs with high accuracy, there are cases where it mispredicts a significant number non-TM helices as TM. For instance, in the bovine mitochondrial ATP synthase state 2a (5ARH), 10 of the 38 predicted TM helices were found to be misclassified (Fig. 5).

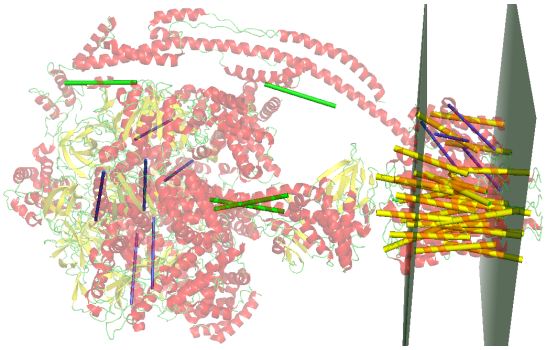

Figure 5: *Results of the helix prediction and membrane calculation of 5ARH. All helices represented as a cylinder were classified as transmembrane by the SVM. Blue vectors were removed during the first filter step, green vectors after the second. The remaining yellow vectors were used for membrane computation.*

These results cannot be avoided when only using amino acid frequencies as features, as non-TM helices with a hydrophobic amino acid composition are known to exist, particularly in the protein core. However, this further reinforces the need to reevaluate potential TM helices using geometric criteria.

It is possible to calculate a membrane geometry using the TMDET algorithm in an exhaustive manner with a large number of test vectors, but calculating the geometry is quite time-costly (ca. 10 s per calculation) and scales with the number of tested vectors. Therefore, calculating a tentative orientation using the predicted TM helices of a protein and only testing vectors in close proximity reduces the search space and computation time drastically.

Considering the removal of false positives during the filtering step, this work can be the basis of a two-step transmembrane helix classifier, which makes predictions based on a combination of the amino acid content of each helix (SVM) and geometrical features of its tertiary structure (filtering). This refined classifier would most likely lead to an even higher specificity.

All source code is available at `https://github.com/spaethju/transmembrane-identifier`.

References

[1] Stephen White. General Principles of Membrane Protein Folding and Stability. `http://blanco.biomol.uci.edu/mp_assembly.html`, accessed 2018-06-18.

[2] B Rost, R Casadio, P Fariselli, C Sander, Burkhard Rost, Rita Casad, Pier Fariselli, and Chris Sander. Transmembrane helices predicted at 95 % accuracy Transmembrane helices predicted at 95 Yo accuracy. *Protein Science*, pages 521–533, 1995.

[3] Anders Krogh, È Larsson, Gunnar Von Heijne, and Erik L L Sonnhammer. Predicting Transmembrane Protein Topology with a Hidden Markov Model : Application to Complete Genomes. *Journal of molecular biology*, 2001.

[4] Gábor E. Tusnády, Zsuzsanna Dosztányi, and István Simon. Transmembrane proteins in the Protein Data Bank: Identification and classification. *Bioinformatics*, 20(17):2964–2972, 2004.

[5] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] D. Eisenberg, E. Schwarz, M. Komarony, and R. Wall. Amino acid scale: Normalized consensus hydrophobicity scale. *Journal of Molecular Biology*, (179):125–142, 1984.