Kajetan Juszczak

Buried versus exposed - transmembrane β barrel residues status prediction

Abstract

Prediction of exposure status of residues in Transmembrane β-barrel is a challenging task. There is only very limited number of β-barrel structures experimentally resolved and therefore the training possibilities are limited by tiny database available. Here support vector classifier(SVC) approach to tackle the problem is presented. The classifier was optimized and further improved by extraction of evolutionary information using PSI-BLAST. The accuracy of this predictor is 74.37% tested on the provided dataset using 3-fold cross-validation, which is comparable to 78,35% achieved by best SVC approaches found in the literature. Also, the predictor is not falling far behind when compared to best available exposure status predictor for β-barrels, namely HMM-based TMBHMM which accuracy is 83%.

Introduction

Support vector machine (SVM) is a supervised machine learning technique. It means that data provided for training is supplied with correct classification. The basic mechanism underpinning the algorithm is a search for the hyperplane which optimally separates training examples and in turn, assigns them to correct classes with the highest possible accuracy. The number of classes have to be defined by the user beforehand(Mountrakis, Im, & Ogole, 2011). One of easily accessible ways to implement SVM is through scikit-learn library(Varoquaux et al., 2015). There are 4 main types of kernels in SVM provided by scikit-learn: linear, polynomial, rbf and sigmoid. These kernel types define the shape of the hyperplane. Different kernels might be favourable to use depending on the underlying problem one is trying to solve.

Assessing the accuracy of the model is important for getting an estimate of its predicting capabilities and also for best model selection. One way of accuracy assessment is cross-validation(Kohavi, 1995). In this technique, the training database is split into parts. some of them are used for algorithm training and some are only used for testing the performance of the model. In case of K-fold cross-validation, the dataset is split into k parts. All parts except one are used for training and remaining part is used for testing. The process is repeated k times, so as every part is used for testing once. This approach allows for generalization the accuracy of the model, as test set is simulating real problem, never seen by model before.

Transmembrane β-barrels are transmembrane proteins formed by antiparallel β-sheet arranged in barrel shape structure. Two following residues within β-strand are always pointing in the opposite directions forming the in/out pattern with reference to the center of the barrel. The residues pointing outside will always be nonpolar, as they are facing nonpolar inner part of the membrane. Residues pointing inside will always be polar(Zvelebil, M., & Baum, 2007). There are very little β-barrels structures experimentally resolved so far. Predicting transmembrane regions is difficult since β-barrels lack characteristic features such as a stretch of 15-30 consecutive hydrophobic residues or positive inside rule present in helical transmembrane protein(Singh, Goodman, Walter, Helms, & Hayat, 2011). Prediction of the exposure status of residues (buried or exposed) is important because of its possible applications in site-specific mutational studies and in channel engineering(Singh et al., 2011).

There are several different approaches for prediction the exposure status of Trans-membrane β-barrel residues. As far as SVM is concerned the best performance in exposure status prediction found in the literature is 77.91% for the membrane core regions and 80.42% in interface regions(Hayat, Park, & Helms, 2011). After further calculation based on data provided in the article, total prediction accuracy for this model is 78,35%. There are some other approaches used to predict the exposure status, as stated by (Singh et al., 2011), the best available one is Hidden Markov Model. TMBHMM which is HMM exposure status predictor achieved a prediction accuracy of 83%(Singh et al., 2011) which is the highest found in the literature.

Methods

- Dataset

The dataset used in this project consisted of 69 transmembrane β barrel non-homologous proteins. For each protein, the exposure status in given position was provided. The dataset was organized in repeating three line pattern: protein ID, protein sequence and exposure status in separate lines for each entry.

- Including evolutionary information – PSI-Blast profiles

In order to add evolutionary information which might improve the accuracy model, PSI-BLAST(Altschul, 1997) was used to generate PSSM for each protein in the dataset. SwissProt database was chosen as a reference database for PSI-BLAST instead of UniRef90 in the interest of time, as it allowed to decrease the time necessary to perform this step drastically. E-value was set to 0.01 and number of iterations to 3. Obtained profiles were stored in a subdirectory as separate files for each protein in the database.

- Extracting features from the dataset

For this purposed 3 separate lists were created one for storing protein ID, one for PSSM profiles and one for exposure status. To each list, a related line from dataset file was appended. List of exposure status had to be converted from strings into SVM input format, in this case, an array of 0 and 1. PSSM profiles had to be transformed first in order to be used in following steps. It was done with np.genfromtxt function saving only frequency matrix as a 2D array where each row was describing the probabilities of each amino acid in this position. The percentage values were stored as fractions to avoid biases. The lists were created in such way that same indexes in each list corresponded to the same protein.

- Creating sliding window and corresponding states.

In order to obtain input format accepted by SVM, an array for each window was created. To avoid confusion, window length had to be an odd number. The length of each array was $20 \cdot windowlength$ as there are 20 numbers describing probabilities of amino acid in given position of sequence. For window length n the window in position(i) consisted of frequency arrays of residues from $i - \frac{n-1}{2}$ to $i + \frac{n-1}{2}$. An important feature which had to be taken into consideration was solving border cases - windows which range was going over the ends of the sequence. In this cases instead of frequency information an array consisting of 20 zeros were added for each position over the range of the sequence. All windows for all proteins were stored together as 2D array with shape:

$$[number\ of\ all\ windows, 20 \cdot windowlength]$$

The corresponding states were appended in such way that the index of the array of states was the same as the index of the window in all windows array.

- Cross-validation and model optimization

In order to obtain the generalized accuracy of the model, 3-fold cross validation was performed using the cross_val_score function from sklearn library(Varoquaux et al., 2015). 3-fold was chosen since it takes significantly less time to run compared to often used 10 fold cross validation. The parameters were tweaked one by one for window lengths between 3 and 21. The range of window lengths was set based on tests as accuracies for window lengths above 21 were generally lower. All possible kernels for SVC(linear, polynomial, rbf and sigmoid) and also LinearSVC were tested. The cache_size parameter was set to 3000 to speed up the process. Finally, the results for two other methods – random forest classifier and simple decision tree were generated for the same range of window length. Model for best scoring SVC parameters was generated and stored using pickle in the model directory as PSSM_model.

- Predictor and results generation

Program for prediction was written in a similar way as modeling one. Provided fasta file with proteins of unknown exposure status, it generates windows, this time however instead of frequency matrix, the sequence is converted into binary form. For each sequence in testing dataset, the exposure status is predicted based on previously generated model and stored in the results directory in the three line pattern. Results of all the optimizations were stored in MS Excel, where later, plots were generated. Confusion matrix, receiver operating characteristic(ROC) curve and Matthews correlation coefficient (MCC) were generated using sklearn library functions(Varoquaux et al., 2015).

Results and discussion

In order to obtain best possible accuracy of the model it is necessary to try different parameters of SVC. In this project, different kernels at different window lengths were tested first. The results of tests are visible in *Figure 1.*

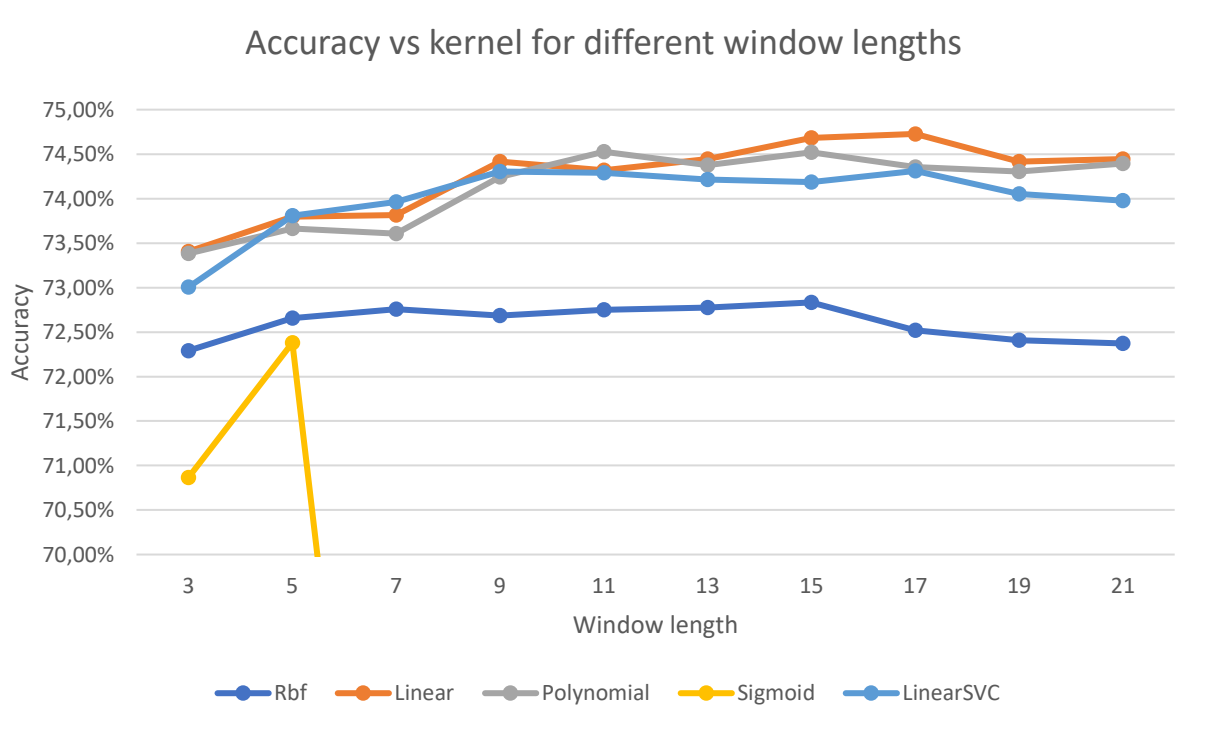

Figure 1 Accuracy of SVC for different kernel types and window lengths

The accuracy values are presented as percentage values, which are the average of scores for 3 fold cross-validation. Linear kernel for SVC was the one which achieved the highest peak value of 74.73% and it was observed for a window length of 17. Polynomial kernel characterized by degree = 4 and coefficient =2 obtained slightly lower results with the highest score being 74.53% for the window length of 11. For polynomial kernel the accuracy scores were rising with a higher degree of kernel, however, this behavior might be credited to overfitting and therefore I decided to pick degree of 4 as the highest value. LinearSVC was characterized by similar results as two abovementioned SVC kernels. Highest Accuracy value was observed for a window length of 17 and it was 74.31%. Rbf kernel marked on the figure with dark blue color had significantly lover accuracy with the highest value being 72.83% for a window length of 15. In case of sigmoid kernel showed with yellow color, not all accuracy values were presented on the figure in the interest of focusing on the more relevant results. For this kernel, the accuracy was rising and peaked 72.38% at window length 5 after which it plummeted to the level of around 51% for a window length of 11, which is in fact almost random prediction, and remained on the level for the rest of tested window lengths. Since the accuracy of SVC with linear kernel was highest, it was taken for further comparison with other methods. For each kernel tested multiple additional tests were performed in order to provide the best possible parameters for model training. It was checked that changing of tolerance value didn't improve the accuracy of models. When C value was manipulated, all changes from default value resulted in same or decreased accuracy scores. the class_weight parameter wasn't changed since the dataset provided was rather balanced: 6528 exposed residues and 6937 buried residues.

Figure 2 presents the accuracies of best SVM model created with multiple sequence alignment inputs and SVM model with the highest score when input was a single sequence.

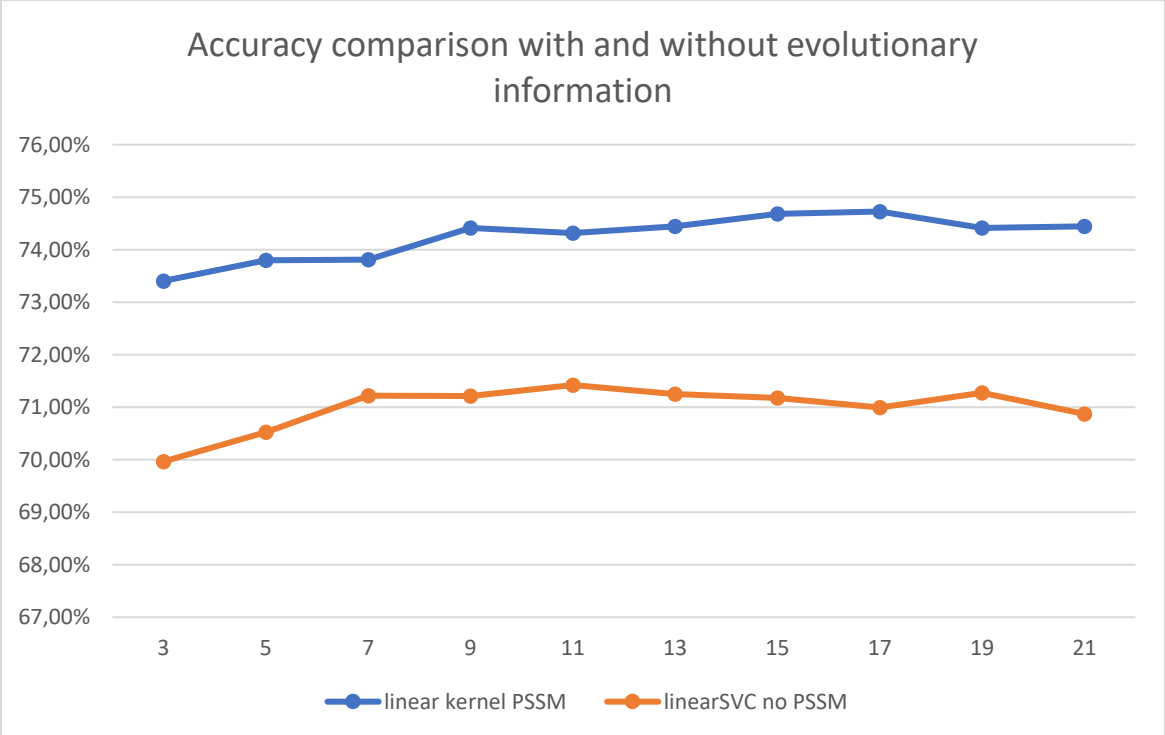

Figure 2 Comparison of the results with and without the addition of evolutionary information

As clearly visible there is a noticeable improvement in prediction accuracy when evolutionary information in for of PSSM frequency matrix was added. The difference between highest accuracies is

more than 3% - 71.42% is highest for single sequence information(window length = 11) compared to 74.73% for highest accuracy when input was multiple sequence alignment data.

The results of comparison between SVC and two other classification techniques, namely random forest classifier(RFC) and decision tree classifier(DTC) are displayed in Figure 3. Both random forest classifier and decision tree classifier were tested on default parameters. The highest accuracies for all window lengths were achieved by RTC, with the highest accuracy being 76.37% for window length equal 5. The results for SVC have already been discussed above, as for other comparisons, SVC with best possible parameters was used.

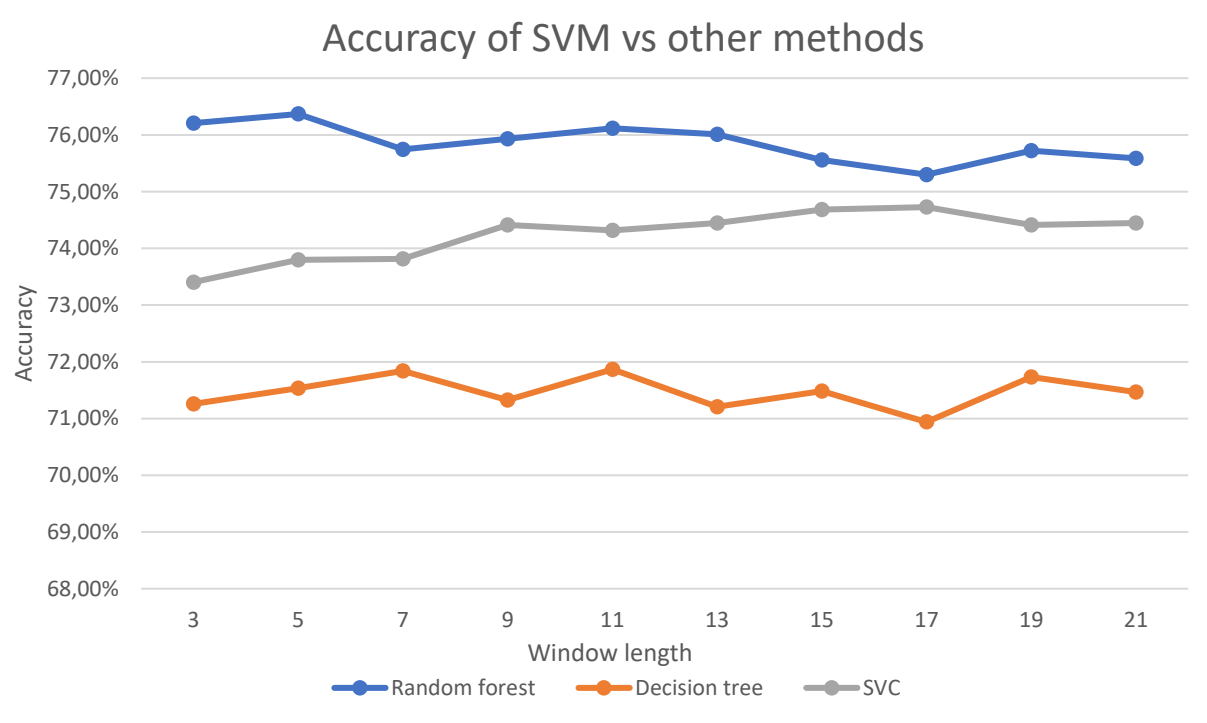

Figure 3 Comparison of best SVC with decision tree and random forest classifier methods

DTC predictions were significantly less accurate and peaked with 71.78% accuracy at window length of 11. For the highest accuracy values achieved by each classifier, Matthews correlation coefficient was calculated, and it is presented in Table 1, supplied with optimal window length for each of classifiers.

Table 1 Matthews correlation coefficients for different classifiers

Classifier	Optimal window length	Matthews correlation coefficient (MCC)
SVC	17	0,496201
RFC	5	0,532415
DTC	11	0,441449

Additional measurements were also performed for the optimal SVC model, such as receiver operating characteristic(ROC) curve and confusion matrix presented in figure 4. There is a slight difference in the percentage of correct predictions between both classes. Looking at figure 4b it can be observed that 72% of all buried residues were predicted correctly, while correct prediction rate was 5% higher in case of exposed residues. The difference is not very big and cannot be attributed to bias towards exposed residues in training dataset since none of classes were overrepresented there.

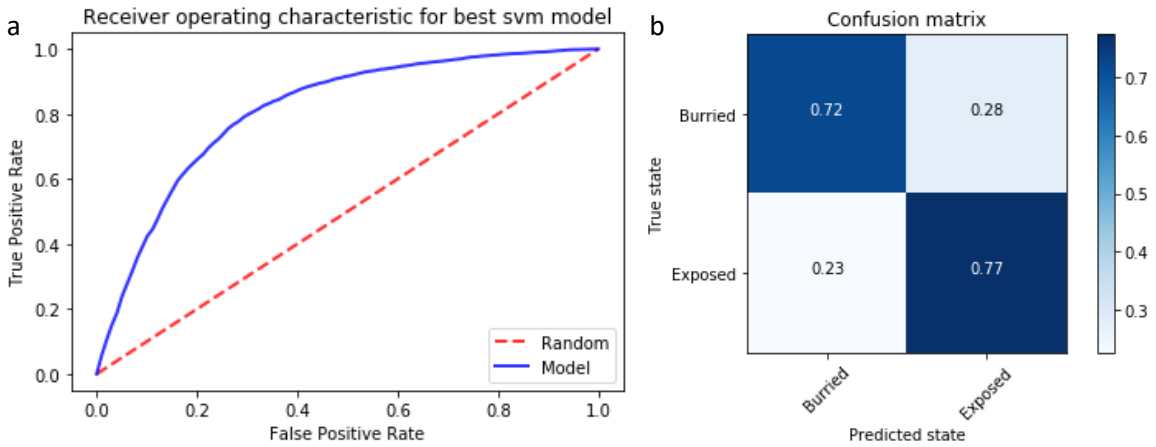

Figure 4a. Receiver operating characteristic curve for a model with optimal SVC parameters. 4b. Confusion matrix for the same model.

Conclusions

To sum up, in this course exposure status predictor for transmembrane β barrels using SVC was developed. Linear kernel achieved the best results, slightly higher than other kernels. Addition of evolutionary information in form of PSSM matrices obtained by PSI-BLAST, significantly(3%) increased the prediction accuracy. The final optimized model was compared with models created using DTC and RFC. The accuracy of SVC was better than DTC but worse than RFC. When accuracy of predictions was compared to the accuracy for SVC approaches found in literature, the results were quite similar. TMBHMM which is considered best exposure status predictor is 8% more accurate compared to this model.

Altschul, S. F. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Hayat, S., Park, Y., & Helms, V. (2011). Statistical analysis and exposure status classification of transmembrane beta barrel residues. *Computational Biology and Chemistry*, *35*(2), 96–107. https://doi.org/10.1016/j.compbiolchem.2011.03.002

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Appears in the International Joint Conference on Articial Intelligence (IJCAI)*, 1–7. https://doi.org/10.1067/mod.2000.109031

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Singh, N. K., Goodman, A., Walter, P., Helms, V., & Hayat, S. (2011). TMBHMM: A frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. *Biochimica et Biophysica Acta - Proteins and Proteomics*, *1814*(5), 664–670. https://doi.org/10.1016/j.bbapap.2011.03.004

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, a. (2015). Scikit-learn: Machine Learning Without Learning the Machinery. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Zvelebil, M., & Baum, J. (2007). *Understanding Bioinformatics*. *Garland Science*.