

Prediction of transmembrane protein crystallization

Julia Varga^{1,2}, László Dobson² and Gábor E. Tusnady²

1) Department of Applied Biotechnology and Food Science, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Budapest, Hungary

2) „Momentum” Membrane Protein Bioinformatics Research Group, Institute of Enzymology, Research Centre for Natural Sciences of the Hungarian Academy of Sciences, Budapest, Hungary;

INTRODUCTION

Transmembrane proteins play vital roles in cellular processes, by taking part in cell-cell adhesion, acting as channels, receptors and enzymes. 25% of the human proteome are transmembrane proteins, targeted by about half of the marketed drugs. However, due to their special physical-chemical properties, only 2% of deposited protein structures belong to transmembrane proteins.

Our aim is to develop a method to **aid the crystallization of transmembrane proteins**. The method consists of two parts: a **database** that identifies transmembrane proteins whose structure would help the modeling of the highest number of transmembrane proteins with yet unknown structures; and a **prediction method** that would help to determine the crystallization propensity of each transmembrane protein.

METHODS

Sequences of transmembrane proteins were extracted from **TargetTrack**¹ database and categorized into positive and negative datasets. Sequences of **PDBTM**² structures determined by X-ray crystallography were categorized as positives. Sequences with >40% identity to sequences in the positive set were removed from the negative dataset. Both datasets were clustered to 40% sequence identity with CD-HIT.

Topology of sequences was determined with CCTOP algorithm³, sequence features were calculated using ProtParam⁴, AAIIndex⁵ properties and in-house scripts.

Support vector machines are being trained, using caret R package, with 10-fold cross-validation and recursive feature elimination to optimize the parameters of training and hyperparameters of the models.

PDBTM

TargetTrack

Dataset	Positive	Negative
Solubilization	1965	1000
Purification	1469	518
Crystallization	486	849

Feature extraction

ProtParam

pI, GRAVY, MW
instability index

AAIndex

amino acid properties

CCTOP

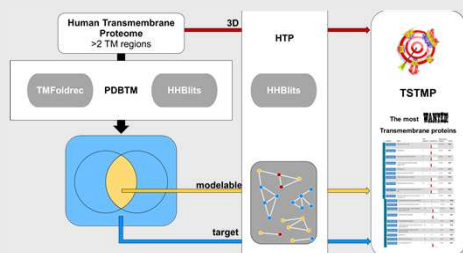
topology

Sequence

length
amino acid composition

PRELIMINARY RESULTS

THE TSTMP DATABASE



In TSTMP database⁶ we examined the **possibility of modeling a transmembrane protein structure** assuming prior knowledge about its relatives, and created a **list of target proteins** from the human transmembrane proteome, as well as a possible strategy to achieve a higher structural coverage.

We also provided the data of proteins that have solved 3D structures covering all transmembrane segments, as well as possible template structures for modelable proteins.

Machine learning

First layer SVMs

Solubilization

Purification

Crystallization

Second layer SVM

The database can be accessed at <http://tstmp.enzim.ttk.mta.hu/>



THE PREDICTION METHOD

A prediction model have been trained for crystallization data so far. The results are promising: by using only a limited number of features, the **model performed quite well**. Possibilities to enhance the model performance should be considered based on these results and literature. These would include adding new features, trying different ways of dealing with unbalanced training data or even training new models.

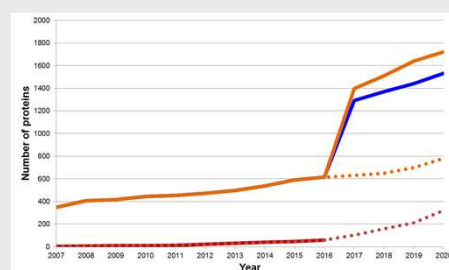
REFERENCES

1. Gabanyi, M. J. et al. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics* **12**, 45–54 (2011).
2. Kozma, D., Simon, I. & Tusnady, G. E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* **41**, D524–D529 (2013).
3. Dobson, L., Reményi, I. and Tusnady, G. E. (2015) CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.*, **43**, W408–W412
4. Wilkins, M. R. et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112**, 531–52 (1999).
5. Kawashima, S., Ogata, H. & Kanehisa, M. AAIIndex: Amino Acid Index database. *Nucleic Acids Res.* **27**, 368–369 (1999).
6. Varga, J., Dobson, L., Reményi, I. & Tusnady, G. E. TSTMP: target selection for structural genomics of human transmembrane proteins. *Nucleic Acids Res.* **45**, D325–D330 (2017).

INTEGRATING DATABASE AND PREDICTION

Around **half of the entries** of the 'Most wanted list' of the database do **not have TargetTrack entry** and a further quarter only have the status 'selected'.

A **modified version of the target list** of the TSTMP database was created by incorporating the developed prediction method. The extrapolated numbers of modelable proteins (blue) are similar to the ones without prediction (orange) and both are considerably higher than the extrapolation of current trends (dashed). Results are based on the extrapolated number of solved 3D structures (red) until 2020.

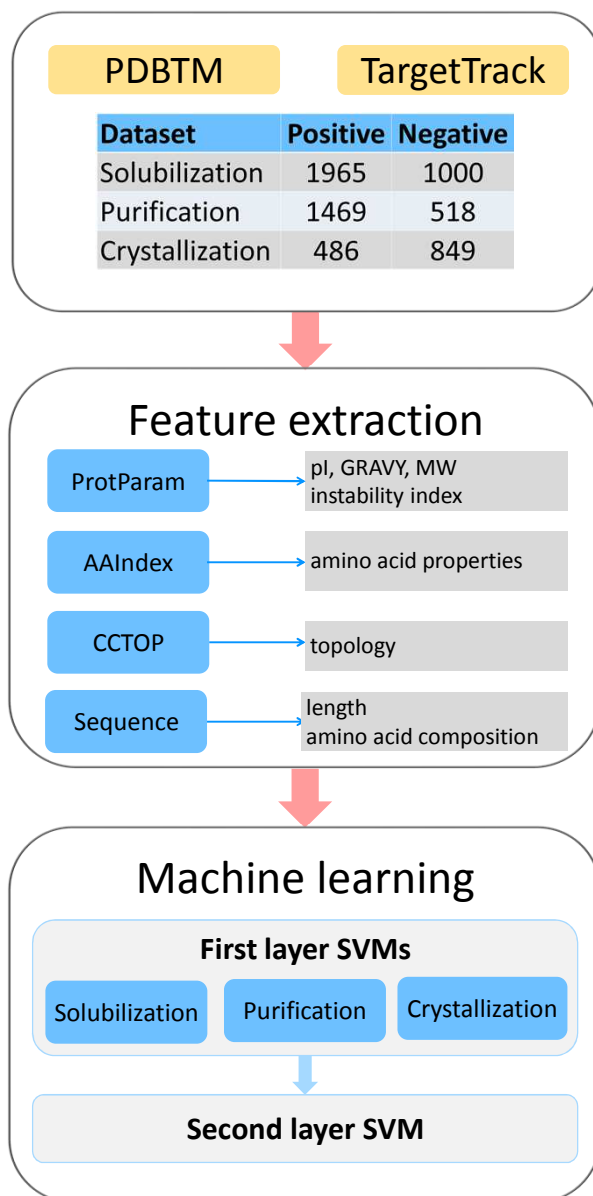


CONCLUSIONS

We established a database to help with the **target selection of human transmembrane proteins** for structure determination, providing a possibility to cover higher proportion of the structural space. A **prediction method** to further aid several steps of the structure determination process is currently under development.

Dataset	Positive	Negative
Solubilization	1965	1000
Purification	1469	518
Crystallization	486	849

Accuracy	0.76
Sensitivity	0.64
Specificity	0.86
Balanced accuray	0.74
MCC	0.51



Accuracy	0.76
Sensitivity	0.64
Specificity	0.86
Balanced accuray	0.74
MCC	0.51

