

## Structural genomics

# The TMCrys server for supporting crystallization of transmembrane proteins

Julia K. Varga and Gábor E. Tusnády\*

"Momentum" Membrane Protein Bioinformatics Research Group, Institute of Enzymology, Research Center of Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest, Magyar tudósok körútja 2.

\*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

### Abstract

**Motivation:** Due to their special properties, the structures of transmembrane proteins are extremely hard to determine. Several methods exist to predict the propensity of successful completion of the structure determination process. However, available predictors incorporate data of any kind of proteins, hence they can hardly differentiate between crystallizable and non-crystallizable membrane proteins.

**Results:** We implemented a web server to simplify running TMCrys prediction method that was developed specifically to separate crystallizable and non-crystallizable membrane proteins.

**Availability:** <http://tmcrys.enzim.ttk.mta.hu>

**Contact:** [tusnady.gabor@ttk.mta.hu](mailto:tusnady.gabor@ttk.mta.hu), [varga.julia@ttk.mta.hu](mailto:varga.julia@ttk.mta.hu)

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Transmembrane proteins (TMP) play vital roles in the cells acting as gatekeepers and receptors in the cell and organelle membranes. They are frequently targeted by pharmaceuticals: a survey found that more than 50% of marketed drugs interact with TMPs (Hopkins and Groom, 2002). Although the human proteome consists of about 25% TMPs (Dobson *et al.*, 2015), however, of all known protein structures only 2% belong to them (Kozma *et al.*, 2013) and less than a hundred human TMP non-redundant structure is determined (Varga *et al.*, 2017). Knowing the structure of TMPs may aid drug development by providing targets for ligand screening and enabling the creation of models for proteins with unknown structures. However, membrane proteins reside in the cell membrane making the process of structure determination extremely difficult.

In the last 10 years, several prediction methods were developed to enhance the success of structure determination by estimating the chance of successful experiments. Most of them uses the data from TargetTrack (Berman *et al.*, 2009) or its predecessors PepcDB and TargetDB (Chen *et al.*, 2004) and PDB structures (Kouranov *et al.*, 2006). However, almost all of them mix globular and TM proteins leading to predict TMPs as 'hard to crystallize' (or somewhat equivalent) without the ability to distinguish between crystallizable and non-crystallizable TMPs. The only TMP-specific method is MEMEX (Martin-Galiano *et al.*, 2008) but being

created in 2008, the data used is outdated. We introduced the TMCrys (Varga and Tusnády, 2018) method to aid the process of structure determination of TMPs. Since the algorithm of TMCrys requires installing some libraries and software packages hereby we introduce the TMCrys server, providing a graphical user interface for the prediction via our HPC to facilitate the usage of the method.

## 2 Methods

### 2.1 Introduction to TMCrys

Training and test datasets for TMCrys were created using PDBTM and TargetTrack databases as described in (Varga and Tusnády, 2018). Several physical and chemical features describing the sequences were calculated using the topology of the protein, predicted by CCTOP algorithm (Dobson *et al.*, 2015), and other programs (Xiao *et al.*, 2015; Overton and Barton, 2006; Petersen *et al.*, 2009; Walker, 2005). Three XGBoost Decision Trees models were trained to predict the success of purification, solubilization and crystallization, respectively. Finally, a model aggregating the results of the three steps were computed to predict the success of the whole process. The models were evaluated using 10-fold cross-validation and tested on their respective hold-out datasets.

© The Author(s) 2019. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

## 2.2 Reliability of the predictions

Reliability of the prediction were defined as the distance from the threshold of the calculated probabilities, normalized to one:

$$reliability = \frac{0.85 + (0.15 * abs(probability - threshold))}{threshold}$$

where threshold for the whole process was 0.85.

## 2.3 The TMCrys server

TMCrys server was developed using the Laravel web application framework (version 5.5.2) and designed with Bootstrap 3.2.7. Upon submitting a job, the sequences are forwarded to a high-performance computing (HPC) cluster. An Apache Axis server monitors the jobs on the cluster and provides the base of the communication between the HPC and the hosting server. The status of the job and the results are retrieved using SOAP requests. Several programs and scripts are run simultaneously to calculate features for the prediction to speed up the process. The results are sent back to the web server and displayed in HTML format and links are available for the download of the results in XML or tab separated format. Users may provide a job name for the identification of their job and optionally an email address as the results usually takes several minutes to obtain. An overview of the prediction process is provided in Supplementary Figure 1.

## 3 RESULTS AND DISCUSSION

### 3.1 Input

The server accepts input in several formats. Basically, one can submit sequences in FASTA format or space separated format. As the topology of the membrane protein is required for calculating the features, the user is permitted to submit topology of the protein calculated by themselves that should have the same length as the sequence and can contain the following labels: 'I' for inside, 'M' for membrane, 'O' for outside, 'L' for re-entrant loops and 'S' for signal peptide. Since the final prediction depends on the topology provided, the user submitted topology might influence the final results. To avoid server overload, maximum 10 sequences can be submitted as one job. The sequences can also be uploaded in a single file.

### 3.2 Output

Three typical HTML outputs can be seen on Supplementary Figure 2. The server generates HTML output for all query proteins in the following format. A query protein appears in an expendable panel. The color of the panel gives information about the protein being membrane or non-membrane, the latter indicated with a yellow panel and 'non-TMP' label (Supplementary Figure 2C). When the protein was predicted to be membrane protein by CCTOP (or a topology was provided), a green or a red panel appears indicating whether the protein was predicted to be crystallizable (Supplementary Figure 2A) or non-crystallizable (Supplementary Figure 2B), respectively.

The predicted outputs are provided in numerical formats as well as a slider diagram, together with the reliability of the prediction. Besides the sequence and the topology of the query, similar entries from TargetTrack and TSTMP databases - generated by simple blast search - are also listed. The former ones aid the process by providing TargetTrack IDs of similar experiments already performed. The TSTMP is a database that collects human membrane proteins with existing structures that can be used for modeling the query protein (group 3D), membrane proteins that can be

modeled (group Modelable) and proteins without existing structure or model (group 'Target'). These latter proteins would become modelable if the structure of the query protein was solved. Last, some of the calculated features are also displayed, like instability index or average solvent accessible surface area.

The outputs can be downloaded in XML and tab-separated format, displaying all the above described features and outputs.

### 3.3 Direct interface

To enable programmatic access to TMCrys server a direct interface was established as well. The user can submit one sequence at a time with an ID and can monitor the progress of the job by calling a polling interface. The results can be downloaded in both tab or XML formats. A template script developed in Python, that can process multiprotein FASTA files, is also provided on the server.

## Acknowledgements

We thank László Dobson for creating the script for accessing the direct interface.

## Funding

This work was supported by the Hungarian Scientific Research Fund [grant number K119287 and K125607]; "Momentum" Program of the Hungarian Academy of Sciences [grant number LP2012/35]; National Research, Development and Innovation Fund of Hungary [grant number FIEK\_16-1-2016-0005] and grant of the New National Excellence Programme by the Ministry of Human Resources [grant number UNKP-16-2\_VBK-016]. Funding for open access charge: LP2012/35.

*Conflict of Interest:* none declared.

## References

- Berman, H.M. et al. (2009) The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.*, **37**, D365-8.
- Chen, L. et al. (2004) TargetDB: A target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860-2862.
- Dobson, L. et al. (2015) The human transmembrane proteome. *Biol. Direct*, **10**, 31.
- Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727-30.
- Kouranov, A. et al. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302-D305.
- Kozma, D. et al. (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524-D529.
- Martin-Galiano, A.J. et al. (2008) Predicting experimental properties of integral membrane proteins by a naive Bayes approach. *Proteins Struct. Funct. Genet.*, **70**, 1243-1256.
- Overton, I.M. and Barton, G.J. (2006) A normalised scale for structural genomics target ranking: The OB-Score. *FEBS Lett.*, **580**, 4005-4009.
- Petersen, B. et al. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.
- Varga, J. et al. (2017) TSTMP: target selection for structural genomics of human transmembrane proteins. *Nucleic Acids Res.*, **45**, D325-D330.
- Varga, J.K. and Tusnády, G.E. (2018) TMCrys: predict propensity of success for transmembrane protein crystallization. *Bioinformatics*.

***TMCryst server***

- Walker, J.M. ed. (2005) *The Proteomics Protocols Handbook* Humana Press, Totowa, NJ.
- Xiao, N. *et al.* (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**, 1857–1859.