Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Application Note

Structural bioinformatics

PredMP: a web server for *de novo* prediction and visualization of membrane proteins

Sheng Wang[1,†,*], Shiyang Fei[3,†], Zongan Wang[4,†], Yu Li[1], Jinbo Xu[5], Feng Zhao[2,*] and Xin Gao[1,*]

[1]Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. [2]Prospect Institute of Fatty Acids and Health, Qingdao University, China. [3]COMPASS, New York, USA. [4]Department of Chemistry, University of Chicago, USA. [5]Toyota Technological Institute at Chicago, USA.

†Contribute equally. *To whom correspondence should be addressed.

Abstract

Summary: PredMP is the first web service, to our knowledge, that aims at *de novo* prediction of the membrane protein (MP) 3D structure followed by the embedding of the MP into the lipid bilayer for visualization. Our approach is based on a high-throughput Deep Transfer Learning (DTL) method that first predicts MP contacts by learning from non-MPs and then predicts the 3D model of the MP using the predicted contacts as distance restraints. This algorithm is derived from our previous Deep Learning (DL) method originally developed for soluble protein contact prediction, which has been officially ranked No. 1 in CASP12. The DTL framework in our approach overcomes the challenge that there are only a limited number of solved MP structures for training the deep learning model. There are three modules in the PredMP server: (a) The DTL framework followed by the contact-assisted folding protocol has already been implemented in RaptorX-Contact, which serves as the key module for 3D model generation; (b) The 1D annotation module, implemented in RaptorX-Property, is used to predict the secondary structure and disordered regions; and (c) the visualization module to display the predicted MPs embedded in the lipid bilayer guided by the predicted transmembrane topology.
Results: Tested on 510 non-redundant MPs, our server predicts correct folds for ~290 MPs, which significantly outperforms existing methods. Tested on a blind and live benchmark CAMEO from Sep 2016 to Jan 2018, PredMP can successfully model all 10 MPs belonging to the hard category.
Availability: PredMP is freely accessed on the web at http://www.predmp.com.
Contact: sheng.wang@kaust.edu.sa, fengzhao21c@163.com or xin.gao@kaust.edu.sa.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Membrane proteins (MPs) are encoded by ~30% genes and have been targeted by ~50% of therapeutic drugs. Compared to non-membrane proteins (non-MPs), the determination of MP structures is challenging in large part due to the difficulty in establishing experimental conditions where the correct conformation of the protein in isolation from its native environment is preserved. Therefore it is important to develop computational methods to predict MP structures from sequence information.

Though homology modeling (or, template-based modeling) works well for many non-MPs (such as soluble proteins), it encounters some difficulties for predicting MPs partially due to lack of sufficient MPs with solved structures. In particular, currently there are only about 510 non-redundant MPs in Protein Data Bank (PDB), which makes homology modeling infeasible for a large portion of MPs. Thus, *de novo* prediction (or, *ab initio* folding) is needed.

So far the most successful *de novo* prediction methods could be categorized into two classes: fragment assembly, e.g., Rosetta (Kim, et al., 2004) and contact-assisted *ab initio* folding, e.g., CoinFold (Wang, et al., 2016). Fragment assembly approach works mostly on some small proteins but most of the multi-pass transmembrane proteins are relatively large in size; contact-assisted approach heavily depends on accurate prediction of protein contacts, which cannot be achieved either by pure co-evolution methods, such as Gremlin (Kamisetty, et al., 2013) or by methods that exploit co-evolution features using shallow neural networks, such as metaPSICOV (Jones, et al., 2014) on proteins without many sequence homologs (Wang, et al., 2017).

Here we present PredMP, a web server that first predicts the MP structure without using any structural templates, and then visualizes the predicted MP model embedded in the lipid bilayer. The key part of PredMP is the 3D modeling module which is implemented in RaptorX-Contact. The underlying algorithm of this module originates from a Deep

Learning (DL) method mainly developed for soluble protein contact prediction (Wang, et al., 2017), which obtained the highest F1 score in the contact prediction category in CASP12 (Wang, et al., 2017). To overcome the insufficient training data for MP contact prediction, we transfer the knowledge learned from non-MPs to MP contact prediction, and thus call such a method Deep Transfer Learning (DTL) (Wang, et al., 2017). Using the predicted contacts as distance restraints, the 3D model of the MP is constructed by the Crystallography & NMR System (CNS) suite (Brunger, et al., 1998).

With the help of predicted transmembrane topology by DeepCNF (Wang, et al., 2016), the 3D model of the query MP is first embedded into the membrane bilayer using a depth- and residue-dependent membrane burial potential (Wang, et al., 2016), and then visualized by a WebGL-based protein viewer.

2 Workflow and implementation

The basic workflow of PredMP is shown in Figure S1. There are three modules in the PredMP server: (i) the 1D annotation module for the prediction of secondary structure and disordered regions by the RaptorX-Property server (Wang, et al., 2016); (ii) the 3D modeling module for *de novo* generating five 3D models of the query MP by the RaptorX-Contact server (Wang, et al., 2016; Wang, et al., 2017; Wang, et al., 2017), and (iii) the visualization module to display the predicted MPs embedded into the lipid bilayer. Below are the major steps of how PredMP works.

Multiple sequence alignment construction. When the amino acid sequence of an MP is submitted by the user, the server first generates the multiple sequence alignment (MSA) to retrieve the sequence homologs from the protein family to which the input MP belongs.

1D annotation module for local structural property prediction. The MSA is utilized to predict two structural properties of an MP, namely the secondary structure elements and the disordered regions. Specifically, these properties are predicted by RaptorX-Property (Wang, et al., 2016).

3D modeling module for de novo generating MP models. This module consists of two parts: (a) contact map prediction, and (b) 3D model construction. For contact map prediction, the MSA is exploited to predict the residue-residue contact map of an MP by a Deep Transfer Learning (DTL) model that learns from non-MPs (Wang, et al., 2017). For 3D model construction, the 3D models of the input MP are constructed by feeding the predicted secondary structures and predicted contacts to the Crystallography & NMR System (CNS) suite (Brunger, et al., 1998). In brief, the predicted secondary structure is converted into distance, angle and h-bond restraints. We also convert the top predicted contacts to distance restraints. Finally, we build 3D structure models using the CNS suite and select top five models according to the CNS energy function (Wang, et al., 2016). The entire approach is implemented in RaptorX-Contact (Wang, et al., 2017).

Visualization module for the display of the embedded MPs. The final step is the visualization of the embedded 3D model of the input MP into the bilayer membrane, which consists of two procedures: (a) transmembrane topology prediction, and (b) MP embedding. For transmembrane topology prediction, we train a machine learning model DeepCNF (Wang, et al., 2016; Wang, et al., 2015) to predict the 9-label transmembrane region at each residue (Section S3). For MP embedding, we use a similar approach as the Positioning of Proteins in Membranes (PPM) method (Lomize, et al., 2006), which calculates rotational and translational positions of the 3D membrane protein model inside the membrane. The membrane potential is obtained from the statistics of a curated training set of non-homologous transmembrane proteins (Wang, et al., 2016).

3 Performances

The underlying DL method (Wang, et al., 2017) has been blindly tested in CASP12 in 2016 and officially ranked first in the category of protein contact prediction (Schaarschmidt, et al., 2017). Tested on a blind and live benchmark CAMEO (Haas, et al., 2013) from Sep 2016 to Jan 2018, the key module RaptorX-Contact in our server PredMP can successfully model all 10 MPs belonging to the hard category (Section S4).

Here we briefly describe the results on all 510 non-redundant MPs with solved structures in PDB (Table S1). According to the CASP official definition (Kryshtafovych, et al., 2018), for each target the predictor could provide five models. If the best TM-score among the five models is larger than 0.5 to the native structure, then we can claim that this target is correctly predicted (Xu and Zhang, 2010). As shown in Table 1, PredMP significantly exceeds the other methods in terms of accuracy of the TM-score for 3D models and accuracy of the top L/5 (L is the protein length) for predicted contacts. For each target in the 510 dataset, we provide a URL to display/download the 3D models generated by PredMP (http://predmp.com/#/detail/5c6oA).

Table 1. The accuracy of 3D models (first three columns: TM-score, the number of models whose TM-score is above a threshold 0.6 and 0.5, respectively) and the accuracy of long/medium range contact prediction (last two columns: Top L/5, where L is the protein length) on all the 510 non-redundant membrane proteins. Note that we define that a contact is short-, medium and long-range when the sequence separation of two residues in a contact falls into [6, 11], [12, 23], and $\geq$ 24 residues, respectively.

Methods	TMscore	#TM>0.5	#TM>0.6	long	med
Gremlin	0.384	122	56	0.40	0.23
metaPSICOV	0.413	147	77	0.49	0.34
PredMP	**0.547**	**298**	**223**	**0.69**	**0.48**

4 Conclusions and discussions

In this work, we introduced PredMP for *de novo* prediction of membrane proteins (MPs). The server not only allows the accurate modeling of the membrane protein 3D structure, but also enables the embedding of the MP into the lipid bilayer. PredMP was calibrated on a blind and live benchmark CAMEO (Haas, et al., 2013) from Sep 2016 to Jan 2018 and successfully modelled 10 MPs. We also constructed a reliable correlation curve between the 3D modelling accuracy and the number of effective sequence homologs (Section S5), and estimated that our server could predict correct folds for ~1,500 among 2,215 human multi-pass MPs including a few hundred new folds (Wang, et al., 2017). This website is free and open to all users and there is no login requirement. The only required input is the putative membrane protein sequence and the running time of our server is about 2 hours per target with about 500 residues. Section S6 details the input/output format of the PredMP server.

We have made available the predicted models as well as the native structures of the 510 non-redundant MP dataset, which is free to access at http://www.predmp.com/#/download. Users can evaluate the quality of the results generated by PredMP. We hope that this 510 dataset could serve as an MP benchmark for the protein prediction community.

References

Brunger, A.T., *et al.* (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination, *Acta Crystallographica-Section D-Biological Crystallography*, **54**, 905-921.

Haas, J., *et al.* (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information, *Database*, **2013**.

Jones, D.T., *et al.* (2014) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins, *Bioinformatics*, **31**, 999-1006.

Kamisetty, H., Ovchinnikov, S. and Baker, D. (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era, *Proceedings of the National Academy of Sciences*, **110**, 15674-15679.

Kim, D.E., Chivian, D. and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server, *Nucleic acids research*, **32**, W526-W531.

Kryshtafovych, A., *et al.* (2018) Assessment of model accuracy estimations in CASP12, *Proteins: Structure, Function, and Bioinformatics*, **86**, 345-360.

Lomize, A.L., *et al.* (2006) Positioning of proteins in membranes: a computational approach, *Protein Science*, **15**, 1318-1333.

Schaarschmidt, J., *et al.* (2017) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age, *Proteins: Structure, Function, and Bioinformatics*.

Wang, S., *et al.* (2016) RaptorX-Property: a web server for protein structure property prediction, *Nucleic acids research*, **44**, W430-W435.

Wang, S., *et al.* (2016) CoinFold: a web server for protein contact prediction and contact-assisted protein folding, *Nucleic acids research*, **44**, W361-W366.

Wang, S., *et al.* (2017) Folding membrane proteins by deep transfer learning, *Cell systems*, **5**, 202-211. e203.

Wang, S., *et al.* (2016) Protein secondary structure prediction using deep convolutional neural fields, *Scientific reports*, **6**, 18962.

Wang, S., *et al.* (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model, *PLoS computational biology*, **13**, e1005324.

Wang, S., Sun, S. and Xu, J. (2017) Analysis of deep learning methods for blind protein contact prediction in CASP12, *Proteins: Structure, Function, and Bioinformatics*.

Wang, S., *et al.* (2015) DeepCNF-D: predicting protein order/disorder regions by weighted deep convolutional neural fields, *International journal of molecular sciences*, **16**, 17315-17330.

Wang, Z., *et al.* (2016) Including H-Bonding in Depth-Dependent Membrane Burial Potentials for Improving Folding Simulations, *Biophysical Journal*, **110**, 58a.

Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score= 0.5?, *Bioinformatics*, **26**, 889-895.