Bioinformatics, 36(8), 2020, 2595–2598
doi: 10.1093/bioinformatics/btz539
Advance Access Publication Date: 10 July 2019
Applications Note

Structural bioinformatics

MemBlob database and server for identifying transmembrane regions using cryo-EM maps

Bianka Farkas [1,2,3,†], Georgina Csizmadia[1,2,†], Eszter Katona[1,4], Gábor E. Tusnády[5] and Tamás Hegedűs[1,2,*]

[1]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest 1094, Hungary, [2]MTA-SE Molecular Biophysics Research Group, Hungarian Academy of Sciences, Budapest 1094, Hungary, [3]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest 1083, Hungary, [4]Faculty of Brain Sciences, University College London, London W1T 7NF, UK and [5]'Momentum' Membrane Protein Bioinformatics Research Group, Institute of Enzymology, RCNS, Hungarian Academy of Sciences, 1117 Budapest, Hungary

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Lenore Cowen

Received on December 20, 2018; revised on May 31, 2019; editorial decision on June 28, 2019; accepted on July 9, 2019

Abstract

Summary: The identification of transmembrane helices in transmembrane proteins is crucial, not only to understand their mechanism of action but also to develop new therapies. While experimental data on the boundaries of membrane-embedded regions are sparse, this information is present in cryo-electron microscopy (cryo-EM) density maps and it has not been utilized yet for determining membrane regions. We developed a computational pipeline, where the inputs of a cryo-EM map, the corresponding atomistic structure, and the potential bilayer orientation determined by TMDET algorithm of a given protein result in an output defining the residues assigned to the bulk water phase, lipid interface and the lipid hydrophobic core. Based on this method, we built a database involving published cryo-EM protein structures and a server to be able to compute this data for newly obtained structures.

Availability and implementation: http://memblob.hegelab.org.

Contact: tamas.hegedus@hegelab.org

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

1 Introduction

Membrane proteins play an important role in many cellular processes and are highly significant drug targets (Santos *et al.*, 2017; Yin and Flynn, 2016). To understand their folding, maturation and function, and to develop new therapies targeting membrane proteins, determination of both high-resolution structures and their transmembrane (TM) region is crucial. NMR has been applied mostly for small regions of TM proteins, e.g. one TM helix with short flanking regions either in the absence or presence of a membrane mimetics (Berman *et al.*, 2000). Nevertheless, membrane interaction sites were usually not directly tested. In the case of crystallography, lipids in a crystal can be identified infrequently, and in most cases may have attached to non-physiological sites. Experiments, where tags were inserted at various positions around putative TM helices and their accessibility was tested, usually have provided low resolution data (Chang *et al.*, 1994; Zagotta *et al.*, 2016).

Due to the difficulties associated with experimental approaches, various *in silico* methods have been developed to determine the TM region. The most popular methods are the TMDET (Tusnády *et al.*, 2004) and the PPM (Lomize *et al.*, 2011) algorithms that were utilized to generate PDBTM (Kozma *et al.*, 2012) and OPM (Lomize *et al.*, 2012) databases, respectively. These methods deliver the membrane definition as a slab with two parallel planes. Another frequently used database, MEMPROTMD (Stansfeld *et al.*, 2015) provides predictions by building a membrane bilayer around the protein using molecular dynamics simulations.

The revolution in cryo-electron microscopy (cryo-EM) not only led to an increasing number of solved TM protein structures but also allowed the investigation of protein structures in the presence of a lipid environment (e.g. micelle, bicelle and nanodisc). Furthermore, the resulting electron microscopy density maps contain information about the membrane embedment of the protein. We developed a pipeline that extracts the edges of the blob that corresponds to the membrane boundary of the targeted TM protein. We built a database for this hidden information of cryo-EM maps with a resolution better than 4 Å and also a web application to allow the analysis of unpublished densities.

2 Materials and methods

Structures determined by cryo-EM at a resolution higher than 4 Å and their corresponding electron microscopy density maps were downloaded from RCSB and EMDB, respectively (July, 2018). TMDET translation matrices were collected either from PDBTM (Kozma *et al.*, 2012) or by submitting the PDB file to TMDET (Tusnady *et al.*, 2005). As it will be seen below, the center of the TMDET predicted bilayer plays an important role in searching the edge of the density corresponding to the lipid environment (membrane blob).

The pipeline was built on and managed by Python scripts (Fig. 1). In the first step, a non-protein density map was created by subtracting the modeled protein density map of the all-atom structure from the EMD density. The theoretical density map was generated using the VMD MDFF package (Trabuco *et al.*, 2009) at a resolution of 6 Å. We have simulated protein density at resolutions 4, 5, 6, and 8 Å with the 6 Å value providing a slightly cleaner membrane blob compared to the 4 and 5 Å values. This was most likely because the density maps contained regions with lower and higher resolution than 4 Å. Before subtraction, the theoretical map was scaled to the EMD map using the ratio of the largest density values in the two maps. After subtraction, the values below the 10% of the maximal density value were set to zero and the resulted MRC map was converted to 3 D points and corresponding density values.

We found that the start of the search for the blob boundaries was simpler from the inside of the membrane than from the opposite direction. To set the origin in the membrane region, we translated the coordinates by the TMDET matrix, which set the (0, 0, 0) into the middle of a predicted bilayer. Then x-y sections of this translated, experimental, non-protein density map were generated at a frequency of $\Delta z = 2$Å and these sections were slivered from 0 to 350° in angle slices of 10°. The density values in each slice were summed resulting in an array of density values for every z/angle pairs. This array was smoothed by a Savitzky–Golay filter in both dimensions with a window size of five and a polynomial order of three. The first minimum values were identified in both positive and negative z directions from $z = 0$ (set by TMDET) and proposed as the boundaries of the membrane blob. The established boundaries in each slice were projected to the all-atom structure to pair atoms with their localization. Surface atoms that were more distant from the bilayer center than the z-coordinate of the boundary of a given slice were considered as water accessible. By defaults, atoms in an interval of 8 Å from the boundaries toward $z = 0$ were defined to be located within the interface region. 8 Å was used at is it widely applied (Callenberg *et al.*, 2012; Marcoline *et al.*, 2015; Pabst *et al.*, 2000), but this value can be set by the user. Atoms closer to the center were considered to belong to the hydrophobic core. To classify atoms as buried or surface-exposed, DSSP (Kabsch and Sander, 1983) was run using the all-atom structure as an input. The output of DSSP, the boundaries in each slice and the atomic coordinates were combined to set the localization of a residue in the B-factor field of the corresponding PDB file. The values of -10, 0, 5, 10 and 15 sign any undefined localization (unknown residues or non-protein molecules), buried residues, surface residues in the hydrophobic core region, surface residues in the water phase and surface residues in the lipid interface, respectively. We also provided a text file for the easy assessment of the TM regions. PDB files were manipulated using the MDAnalysis Python package (Michaud-Agrawal *et al.*, 2011), while structural images were created using PyMol (The PyMOL Molecular Graphics System, Version 1.8.4 Schrödinger, LLC) and Chimera (Pettersen *et al.*, 2004). The plots with cross-sections and summed densities were generated by Python Matplotlib (Hunter, 2007).

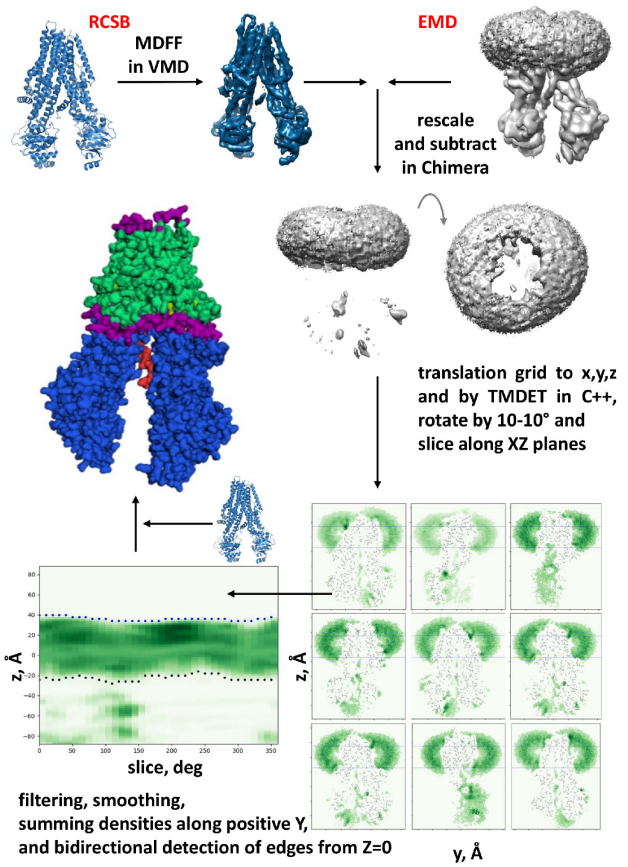

Fig. 1. Main steps of membrane region determination. The density of the protein calculated from the atomistic structure is subtracted from the whole density [CFTR, PDBID: 5UAK (Liu *et al.*, 2017)]. The remaining density is smoothed and projected to 2D. The boundaries of slices are determined from this matrix and mapped back to the all-atom structure

A database and a web application were created to make the pipeline and the results of our runs available. MariaDB (http://mariadb.org) is used as a database backend to store the parameters of submitted jobs and the calculated data of proteins with a resolution better than 4 Å. SQLAlchemy (http://www.sqlalchemy.org) is employed for the object/relational mapping and TurboGears web framework (http://www.turbogears.org) to tie the data, logic and presentation layers into a web application. The main calculation runs independently from the web application in a linear queue system and completed in 1–2 min, which is reasonable, considering the publication frequency of new experimental membrane protein structures.

3Dmol.js is used to visualize the 3D structure in the web page (Rego and Koes, 2015).

3 Results

The web application provides a graphical interface for submitting files and browsing the results, requiring the electron density file in MRC formant and the corresponding all-atom structure in PDB format as inputs. To translate the system and set the (0, 0, 0) coordinate inside the hydrophobic membrane region, the TMDET XML file generated based on the all-atom PDB file is also required. However, if this XML file is not provided by the user, the first four characters of the PDB file name will be treated as a PDBID and used to retrieve the XML file from PDBTM (Kozma *et al.*, 2012). If this process is unsuccessful, the PDB file is submitted to TMDET to obtain the required XML file (Tusnady *et al.*, 2005). On the submit page, the recalculation of the results from our dataset can be initiated by typing a PDB or EMD ID in the appropriate box.

The result page can be accessed upon submitting a new calculation, initiating the recalculation of existing results, and from the browse page of our web application. The result page includes images of *y*-*z* cross-sections at 0, 90, 180 and 270°, while the images of cross-sections at every 10° are packed for download. The plot of the summed and smoothed densities (Fig. 1) is placed into this page as well. The determined boundaries are indicated by blue and black circles. Boundary values outside of the ±1.5 interquartile distance calculated from all *end0* or *end1* boundary values are labeled by triangles. We also put an interactive structural model for visualization on this page to help to decide whether the automatically determined boundaries need manual adjustment. To aid the manual correction of edge detection, we implemented a set of simple commands combined with selection expressions. The three main commands are: (i) *slice_def* end_i ***around*** *z dz*, where *slice_def* is an integer corresponding to a given slice from 0 to 350°, end_i is end0 or end1, *z* is the z-coordinate to search around in the range of (*z-dz, z+dz)*. Slice definition can include comma separated and hyphen separated list of slices (e.g. 10, 40, 100–160) and also an asterisk for all slices. (ii) *slice_def* end_i ***average*** *slice_A slice_B*, which sets the edge of selected slices to the average value of slice A and slice B. (iii) *slice_def* end_i ***equal*** *slice_A*, which set the boundary of the selected slices to the value of slice A's boundary. All the commands are listed in Supplementary Table S1.

We ran the MemBlob pipeline on 92 TM protein structures determined by cryo-EM with a resolution of 4 Å or better. The calculations revealed that approximately 30% of the maps did not exhibit well-defined densities corresponding to the membrane environment (Supplementary Fig. S1). These structures have either been solved in the absence of a well-formed lipid environment or their electron microscopy density maps exhibited a very low signal to noise ratio preventing the detection of the membrane blob boundaries. A good signal to noise ratio of an experimental map is crucial to detect the membrane environment, since densities arising from lipids are significantly lower than those from proteins. In the case of a cation channel [PDBID: 5H3O, (Li *et al.*, 2017)] our pipeline detected rational TM regions in spite of the lack of a membrane blob. A closer look at the density map suggests that the amphipol environment in this case does not contribute to the cryo-EM density, but densities can be observed between the TM helices of the protein. These inter-helical densities indicate the presence of intercalated lipid molecules. We did not find any correlation between the visibility or other properties of the lipid environment and the type of the membrane mimetics (e.g. micelle, nanodisc and amphipol). For example, while the amphipol blob did not contribute to the cryo-EM density in the case of PDBID: 5H3O (Li *et al.*, 2017), it was visible in other instances, such as PDBID: 3J5P (Liao *et al.*, 2013). A summary of the runs is collected in Supplementary Table S2.

We compared the TM region definitions of our pipeline to TMDET predictions, since this *in silico* method has been indicated to provide more feasible boundaries compared to OPM (Koehler Leman *et al.*, 2017). We used it to guide the boundary search in our pipeline. In addition, we have not detected large differences in predictors when previously used for ABC proteins (Csizmadia *et al.*, 2018). For the comparison, first, to get the thickness of the hydrophobic core, we calculated the distance between the boundaries decreased by the thickness of the two interface regions (2 × 8 Å). Then, we averaged the *z*-coordinates of the boundaries for each of the sides resulting in a slab, similar to the output of *in silico* predictors (Supplementary Fig. S2). The membrane center of the MemBlob slab differs from the TMDET center by more than 5 Å only in four cases. In contrast, the MemBlob pipeline determines a thicker membrane environment compared to TMDET. This is often caused by the deep embedment of the protein into the lipids. As a consequence, the location of short regions, which have been considered extra- or intracellular by *in silico* predictors, is indicated intramembranous by our pipeline (Supplementary Fig. S3). This type of membrane-embedment, when the extracellular parts are located in a pit of the membrane, cannot be predicted by *in silico* methods. The physiological role of this embedment may be to provide better protection of the protein from extracellular effects, such as proteases.

4 Conclusions

Most of the cryo-EM studies focus on the determination and characterization of protein structures. However, density maps may contain valuable information other than the well-defined protein density, which has not been fully utilized yet. For example, electron densities derived from disordered protein segments are difficult to extract and interpret. Recently, a machine learning algorithm has been developed for automatic identification of density blobs of ligands from experimental electron microscopy density maps (Kowiel *et al.*, 2019). However, our pipeline is the first that allows the assessment of membrane localization of TM proteins from experimental data at a large scale, using cryo-EM densities. While learning algorithms may supersede the semi-automatic refinement of the boundaries in our pipeline, as of now, we cannot exploit an automatic detection method at this moment due to the low number of cryo-EM maps with sufficient membrane environment densities. Our pipeline possesses two major differences when compared to other existing methods providing TM region prediction. First, MemBlob is fully based on experimental data. While CCTOP supplements its prediction with a large amount of information from experiments (Dobson *et al.*, 2015), this data is coarse-grained (e.g. accessibility experiments), which helps the identification of extramembranous regions rather than the exact location of the bilayer boundaries. Second, MemBlob presents the membrane region as a volume with boundaries that follows the shape of the lipid environment, and not as a slab with parallel edges. MEMPROTMD provides a more realistic configuration of the membrane around the protein using molecular dynamics simulations compared to slab models, but it does not incorporate experimental data other than protein structures (Stansfeld *et al.*, 2015). Therefore, the MemBlob pipeline will be useful for researchers working on structure determination of membrane

proteins using cryo-EM and also for developers of membrane region predictors, who can apply MemBlob results as a true positive experimental set. Since the number of membrane protein structures are expected to rise, the output of our methods will most likely be the starting point to develop automatic methods for the identification of the membrane environment in density maps.

Acknowledgements

Thanks for the suggestions and critical notes to H. Tordai (Department of Biophysics and Radiation Biology, Semmelweis University). We acknowledge NIIF National Information Infrastructure Development Institute (http://www.niif.hu/en) and MTA Wigner GPU Laboratory (http://gpu.wigner.mta.hu) for awarding us access to resources based in Hungary, and the support of their staff is gratefully acknowledged.

Funding

This work was supported by the National Research, Development and Innovation Office [K111678, K119287, K125607, K127961], the Cystic Fibrosis Foundation [CFF HEGEDU18I0] and the Semmelweis Science and Innovation Fund.

Conflict of Interest: none declared.

References

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Callenberg,K.M. *et al*. (2012) Membrane bending is critical for the stability of voltage sensor segments in the membrane. *J. Gen. Physiol*., **140**, 55–68.

Chang,X.B. *et al*. (1994) Mapping of cystic fibrosis transmembrane conductance regulator membrane topology by glycosylation site insertion. *J. Biol. Chem*., **269**, 18572–18575.

Csizmadia,G. *et al*. (2018) Quantitative comparison of ABC membrane protein type I exporter structures in a standardized way. *Comput. Struct. Biotechnol. J*., **16**, 396–403.

Dobson,L. *et al*. (2015) CCTOP: a consensus constrained TOPology prediction web server. *Nucleic Acids Res*., **43**, W408–W412.

Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng*., **9**, 90–95.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–637.

Koehler Leman,J. *et al*. (2017) Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP. *BMC Bioinformatics*, **18**, 115.

Kowiel,M. *et al*. (2019) Automatic recognition of ligands in electron density by machine learning. *Bioinformatics*, **35**, 452–461.

Kozma,D. *et al*. (2012) PDBTM: protein data bank of transmembrane proteins after 8 years. *Nucleic Acids Res*., **41**, D524–D529.

Li,M. *et al*. (2017) Structure of a eukaryotic cyclic-nucleotide-gated channel. *Nature*, **542**, 60–65.

Liao,M. *et al*. (2013) Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature*, **504**, 107–112.

Liu,F. *et al*. (2017) Molecular structure of the human CFTR ion channel. *Cell*, **169**, 85–95.

Lomize,A.L. *et al*. (2011) Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *J. Chem. Inf. Model*., **51**, 930–946.

Lomize,M.A. *et al*. (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*., **40**, D370–D376.

Marcoline,F.V. *et al*. (2015) Membrane protein properties revealed through data-rich electrostatics calculations. *Structure*, **23**, 1526–1537.

Michaud-Agrawal,N. *et al*. (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem*., **32**, 2319–2327.

Pabst,G. *et al*. (2000) Structural information from multilamellar liposomes at full hydration: full *q*-range fitting with high quality x-ray data. *Phys. Rev. E*, **62**, 4000–4009.

Pettersen,E.F. *et al*. (2004) UCSF chimera? A visualization system for exploratory research and analysis. *J. Comput. Chem*., **25**, 1605–1612.

Rego,N. and Koes,D. (2015) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, **31**, 1322–1324.

Santos,R. *et al*. (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov*., **16**, 19–34.

Stansfeld,P.J. *et al*. (2015) MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. *Structure*, **23**, 1350–1361.

Trabuco,L.G. *et al*. (2009) Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, **49**, 174–180.

Tusnady,G.E. *et al*. (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.

Tusnády,G.E. *et al*. (2004) Transmembrane proteins in the protein data bank: identification and classification. *Bioinformatics*, **20**, 2964–2972.

Yin,H. and Flynn,A.D. (2016) Drugging membrane protein interactions. *Annu. Rev. Biomed. Eng*., **18**, 51–76.

Zagotta,W.N. *et al*. (2016) Measuring distances between TRPV1 and the plasma membrane using a noncanonical amino acid and transition metal ion FRET. *J. Gen. Physiol*., **147**, 201–216.