Determination of the topology of the human transmembrane proteome by combined experimental and theoretical methods

PhD Thesis Booklet

Tamás Langó

Structural Biochemistry Program, Doctoral School of Biology
Eötvös Loránd University, Faculty of Science

Head of the doctoral school: Prof. Anna Erdei, DSc
Program leader: Prof. Mihály Kovács, DSc

Supervisor:
Dr. Gábor Tusnády, PhD
Membrane Protein Bioinformatics Research Group, Institute of Enzymology
Hungarian Academy of Sciences, Research Center for Natural Sciences

Budapest
2018
1 Introduction

In the living organism all cells and their organelles are separated by lipid bilayer from the outside environment. Transmembrane (TM) proteins are embedded in the membrane and their polypeptide chain spans it at least once.

![Schematic representation of the topology of an α-helical TM protein](http://topdb.enzim.hu)

Figure 1: Schematic representation of the topology of an α-helical TM protein (blue: extra-cytosolic parts, yellow: TM segments, red: cytosolic parts, Source: http://topdb.enzim.hu)

TM proteins act as molecular “gate” in biological membranes: they play role in different signal transduction and metabolic processes, as well as in maintaining the dynamic equilibrium between the internal and external environment of cells. Treatment for many diseases involves these gates, as they can be used to deliver the active substances. About 55% of the drugs currently approved by the Food and Drug Administration (FDA) target TM proteins. To develop new and effective drugs, it is crucial to deepen our knowledge of their structure and function.

According to the most recent studies, about 25% of the ORFs encode TM proteins. Despite their importance and abundance, only a handful of structures have been determined so far, due to their special physico-chemical properties. Given the persisting technical difficulties associated with high resolution structure determination, bioinformatics and other experimental approaches are widely used to characterize their low-resolution structure.

Topology is the most frequently used representation of TM protein structure, defining the number and location of TM segments, as well as their orientation relative to the membrane (Figure 1). In the recent decades several experimental methods and bioinformatics tools have been developed to determine the topology of TM proteins. The accuracy of some topology
prediction algorithms can be significantly increased by incorporating reliable experimental data as a constraint.

Determination of TM protein topology can be used as a starting point for modeling their 3D structures or laboratory experiments (e.g. revealing accessible regions for antibody or small molecule drug design).
2 Scientific aims

The number of solved TM protein structures is relatively low compared to those of globular proteins. Novel experimental and theoretical methods are needed to reveal their low resolution structure, which can be used to help the various experimental and modeling efforts.

Topology prediction methods can be universally used to gain insights into all TM protein structure, however their result often different compared to experimental data. Some of the prediction methods are able to incorporate previously established topology data, and provide a more accurate prediction. Therefore, the primary purpose of my work was the following:

- Collecting experimental methods from scientific literature, that can produce topology data (such as fusion proteins, post-translational modifications, immunolocalization, protease protection assay etc.).
- Organizing gathered methods and their result to update the TOPDB database.

Most of the currently existing techniques provide limited topology information, as they give information about single proteins. These procedures are often time-consuming and the interpretation of the results are often not straightforward, especially when the localization or function of the protein is changed. Since for more than 50% of the human TM proteome no topology data was available, the secondary goal of my work was as following:

- Developing a new experimental topology assay (MSTOP), that is rapid and produces clear topology data for hundreds TM proteins at the same time. The specific aims of the analytical method development were as following:
  o During early stages of the method setting optimal parameters on surface proteins of red blood cells
  o Development and application of control experiments for each phase of the MSTOP method
  o Optimization of each step of the process, to increase the number of identified TM proteins at the same time.
  o Expanding the MSTOP method for cell lines (HL60, K562).
  o Characterizing sequential environments of the labeled lysine residues.
2 Scientific aims

- Characterizing the accuracy of the MSTOP determined topology data (comparing results with previous experimental data).
- Determining the impact of the method on constrained topology prediction.
3 Methods


Used cells, materials, methods and instruments to MSTOP technique were as following:

- Cells were used: Red blood cell, HL60, K562 and CHO.
- Accessible primer amines were labeled on the cell surface by Sulfo-NHS-SS-biotin reagent.
- Biotinylation of cell surface and membrane impermeability of labeling agent were examined by using FITC conjugated anti-biotin antibody or TexasRed conjugated streptavidin. Fluorescence intensity of fluorochromes was quantified by FACS Attune® Acoustic Focusing Cytometer, and their localization were analyzed by Zeiss LSCM 710 confocal microscopy in cells.
- Cells were lysed in hypotonic lysis buffer, membrane preparations were made by using Beckmann ultracentrifuge L7-55. Protein content was determined by Lowry-method. Quality of preparations was examined by SDS-PAGE gel electrophoresis, and their biotinylation efficiency was analyzed by Western-blot or dot-blot techniques using HRP conjugated avidin.
- Membrane preparations were solubilized (in the presence of Rapigest, iodoacetamide and 2,2’-thiodiethanol), then the suspension was incubated with PNGaseF for 2 hours and the proteolytic enzymes for 16 hours. Digestion was stopped, then the labeled peptides were enriched on neutravidin agarose beads, after that unlabeled peptides were removed by extensive washing steps and the labeled peptides were eluted by reducing agents (DTT or TCEP).
- Eluted peptides were alkylated (by iodoacetamide) and were cleaned on C18 columns. Peptide mixtures were analyzed by LC/MS/MS using two different instrument setups and their software: 1. nanoAcquity UPLC system, LTQ-Orbitrap Elite mass spectrometer, Proteome Discoverer v1.4, Pava script, ProteinProspector v5.14.1, 2. Dionex Ultimate 3000 NanoLC system, CapativeSpray nanoBooster ionization source, Bruker Maxis II ETD Q-TOF mass spectrometer, Compass DataAnalysis 4.3, ProteinScape 3.1, Mascot.
4 Results

4.1 Collecting topology data
Experimental topology information about TM proteins is limited to a few thousand proteins, moreover this data is scattered in the literature. Topology data with sequential positions were gathered from several resources, including literature using PubMed, protein sequence databases such as UniProt, the PDBTM structure database. The processed resources included unique protein assays (such as structure determination of TM proteins, molecular biology techniques characterizing the orientation of TM segments and connecting loops) and high throughput experiments, labeling several proteins at the same time (e.g. proteomics researches determining the sequential positions of N- or O-glycosylations).

The most important unique protein characterizing techniques were collected into the following groups: 3D structure determination, examination of fusion proteins (target protein was fusioned with reporter enzymes or fluorescent proteins), determination of post-translational modifications, experiments based on proteolytic enzymes, immunolocalization of TM protein regions, and specific chemical modification of amino acids.

4.2 Updating the TOPDB database
The TOPDB database was updated using the collected data. Both the number of included TM proteins and the number of experimental data increased by threefold compared to the previous revision. The database is available to the public at http://topdb.enzim.hu. To update the database – after an initial filtering step – thousands of articles were processed manually. Existing TOPDB entries were supplemented with new topology data, while identifiers were generated for newly incorporated proteins.
Figure 2: An example protein from TOPDB database (Folate transporter 1, RFC, UniProt ID: P41440) and its collected topology data. The „Topology” line shows the predicted topology of the protein, the rows below indicate the results of the independent experiments from the related publications (blue: extra-cytoplasmic parts, yellow: TM segments, red: cytosolic parts, Source: http://topdb.enzim.hu).

After updating the database, the greatest increase of assigned data can be noticed in the case of the post-translational modifications, due to the incorporation of high-throughput proteomics methods, identifying dozens of N- or O-glycosylation sites. On the other hand, recently revealed 3D structures provided the highest number of topology data. Altogether 1438 TOPDB entries, containing one or more experimental topology data was collected from different molecular biology techniques. We used the CCTOP prediction algorithm to determine the topology, a novel approach taking into account all the collected experimental topology data during the prediction.

The TOPDB is currently the most complete and comprehensive collection of TM protein datasets containing experimentally derived topology information.
4.3 **MSTOP method**

Although collecting and organizing topology data yielded an enormous amount of topology information, we could not assign any experiment for ~50% of the human TM proteome. To overcome this and to produce experimental data for some of the uncharacterized proteins, we developed the MSTOP method. The proposed approach can produce experimental topology data for hundreds of native TM proteins at the same time (Figure 3).

![Flowchart of the MSTOP method](image)

Figure 3: Flowchart of the MSTOP method, that identifies extra-cytosolic accessible primer amines to produce topology data. We label isolated cells with a membrane-impermeable, primer amine specific labeling agent (sulfo-NHS-SS-biotin). We lyse the cells and purify the membrane fraction, then we solubilize and digest it with different proteolytic enzymes. We isolate the modified peptides on a neutravidin agarose resin, then elute and sequence them by tandem mass spectrometry. The labeled positions we can use as extra-cytosolic constraints in the CCTOP topology prediction algorithm.

The maximum efficiency of cell surface labeling by Sulfo-NHS-SS-biotin reagent was determined by flow cytometry. To verify that the optimal amount of biotinylation reagent has no effect to the integrity of the cells, we monitored the uptake of propidium iodide dye. Successful labeling of cell surface and membrane impermeability of labeling agent were examined by confocal microscopy.

We prepared a membrane fraction from the labeled cells and determined their protein content. We used SDS-PAGE to test the reproducibility of the preparation process, using different samples produced at different times, but with the same protein content. We
confirmed the biotinylation of protein lanes by Western blot, so that biotinylated preparations were loaded onto the gel without reducing agent. Following the solubilization of the preparations, we verified the successful protease digestion using gel electrophoresis.

We also determined the sufficient amount of avidin agarose resin required to bind the total biotin content of each biotinylated membrane preparation using dot blot method. Most of the non-specific peptides were eliminated by adjusting the number of washing steps and buffer composition on the biotinylated peptides carrying avidin resin (to detect biotin content of each washing fraction dot bot method was used as well).

Using the last two steps, we significantly reduced the complexity of the samples, compared to previously published protocols isolating whole proteins with the labeling agent. Instead of labeled proteins, we purified labeled peptides, resulting in a lower amount of identified peptides/proteins, but also ensuring low false discovery rate. As a final step we analyzed produced peptide mixtures using LC/MS/MS utilizing two different instrument setups. The expected artificial modifications were found by both search engines.

The developed MSTOP method was tested on three cell types: on red blood cell, on HL60 and on K562 cell lines. At this point the experiments yielded 730 extra-cytosolic positions for nearly 200 TM proteins. By considering homologous TM proteins, our experiments provided at least one topology data for 2776 human TM proteins in the UniProt database.

To investigate local factors, that may influence successful labeling and mass spectrometry identification, and we analyzed the amino acid composition of the sequential environment (±10 amino acids length) of the labeled position. Based on the analysis our method has a preference to identify labeled positions followed by positively charged amino acids. The lowest amino acid diversity can be noticed in the direct proximity of the labeled positions (at position ±1), suggesting that the neighbouring residues may alter the chemical reaction of the labeling process. We also calculated the length distributions of extra-cytosolic segments, that were identified by covalently modified lysines and compared them to the distribution all connecting loops. The two distributions were not significantly different, thus the length of the extra-cytosolic loops or domains does not affect the reactivity with Sulfo-NHS-SS-biotin.
4.4 Validation of the method and contribution of the results to the CCTOP prediction

To validate our method, we compared the labeled extra-cytosolic lysine residues to independent experimental topology results reported in the literature (TOPDB database). 85 proteins did not have any assigned experimental topology data, however for 113 proteins we found previously determined information. In these proteins extra-cytosolic localization of 450 out of the 730 labeled lysines could be confirmed using by independent experiments. 98.7% of the 450 positions were in agreement with previously established topology data.

To analyze the limitation of the developed method, we investigated the effect of using all extra-cytosolic lysines as constraint during topology prediction. According to the results the maximal benefit is 23% increase in prediction accuracy (from 56% to 79%). By limiting the constraints to 20% of all extra-cytosolic lysines (the capacity of MSTOP), the accuracy of the topology predictions is still increased by 14% (from 56% to 70%). Further calculations demonstrated, that even 4% noise (i.e. false positive, cytosolic lysines used as a extra-cytosolic constraint) reduced the accuracy of topology prediction. According to these calculations and the comparison with independent experimental results, topology data provided by MSTOP can increase the accuracy of topology prediction methods by 14%.
5 Publications

My dissertation is based on the following publications:


Publication is not related to my dissertation: