BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 19 2010, pages 2490–2492
doi:10.1093/bioinformatics/btq362

Databases and ontologies
Advance Access publication July 3, 2010

ExTopoDB: a database of experimentally derived topological models of transmembrane proteins

Georgios N. Tsaousis[1], Konstantinos D. Tsirigos[1], Xanthi D. Andrianou[1],
Theodore D. Liakopoulos[2], Pantelis G. Bagos[2] and Stavros J. Hamodrakas[1,*]

[1]Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis Athens 15701 and [2]Department of Computer Science and Biomedical Informatics, University of Central Greece, Papasiopoulou 2-4 Lamia 35100, Greece

Associate Editor: Martin Bishop

ABSTRACT

Summary: ExTopoDB is a publicly accessible database of experimentally derived topological models of transmembrane proteins. It contains information collected from studies in the literature that report the use of biochemical methods for the determination of the topology of α-helical transmembrane proteins. Transmembrane protein topology is highly important in order to understand their function and ExTopoDB provides an up to date, complete and comprehensive dataset of experimentally determined topologies of α-helical transmembrane proteins. Topological information is combined with transmembrane topology prediction resulting in more reliable topological models.

Availability: http://bioinformatics.biol.uoa.gr/ExTopoDB

Contact: shamodr@biol.uoa.gr

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

Received on April 24, 2010; revised on June 11 2010; accepted on July 1, 2010

1 INTRODUCTION

Transmembrane proteins constitute ∼20–30% of fully sequenced proteomes and they form an important class of proteins, as they are crucial for a wide variety of cellular functions (Krogh *et al.*, 2001). In order to understand their function we must acquire knowledge about their structure and topology across the membrane. However, due to the difficulty of obtaining crystals of transmembrane proteins suitable for crystallographic analyses, a small number of TM proteins have known 3D structure, comprising about 1.5% of all proteins with known 3D structure according to PDB (Berman *et al.*, 2000). Therefore, biochemical and molecular biology methods are routinely used to determine the topology of transmembrane proteins. By topology, we refer to the knowledge of the number and the exact localization of transmembrane segments, as well as their orientation with respect to the lipid bilayer. We have to note however that even in the case of known 3D structures, the exact boundaries of the transmembrane segments cannot be determined accurately (Lee, 2003). Biochemical methods include: techniques of gene fusions, using enzymes such as alkaline phosphatase, β-galactosidase, β-lactamase and various fluorescent proteins, detection of post-translational modifications such as glycosylation, phosphorylation and biotinylation, cysteine-scanning mutagenesis, proteolysis methods and epitope mapping techniques (van Geest and Lolkema, 2000). Prior knowledge produced by such techniques is combined with topology prediction methods resulting in improved prediction performance (Bagos *et al.*, 2006; Melen *et al.*, 2003). Although widely used prediction algorithms are based mainly on machine-learning techniques, mostly Hidden Markov Models (Kall *et al.*, 2004; Krogh *et al.*, 2001; Tusnady and Simon, 2001), during the last few years *ab initio* topology prediction has been shown to be an attainable goal since it yields comparable performance (Bernsel *et al.*, 2008; Hessa *et al.*, 2007).

Although a large amount of information in the literature refers to the determination of the topology of transmembrane proteins, few efforts have been made in terms of collection of this data in a database (Ikeda *et al.*, 2003; Jayasinghe *et al.*, 2001; Lomize *et al.*, 2006; Moller *et al.*, 2000; Tusnady *et al.*, 2005, 2008). Most of these databases are out of date and include small sets of proteins with topologies determined either by X-ray crystallography (3D) or by a combination of various biochemical methods. Here, we report the construction of a database (ExTopoDB), which provides detailed, experimentally verified, annotation of the topology of α-helical transmembrane proteins. In addition, we provide an analysis showing the increase in the prediction performance using topological data from ExTopoDB.

2 METHODS

In order to collect all available information about the topology of α-helical transmembrane proteins we performed an extensive literature search (up to October 2009) in order to find studies that reported the use of biochemical methods for the determination of the topology of transmembrane proteins. We searched PubMed using keywords such as ‘transmembrane’, ‘membrane’, ‘integral’, ‘gene fusion’, ‘phosphorylation’, ‘cysteine scanning’, ‘glycosylation’, ‘biotinylation’, ‘proteolysis’ and ‘epitope mapping’ combined every time with the term ‘topology’. Results were manually filtered as they contained a large number of false positive hits in our search. Topology information was manually retrieved from 1200 studies. When a study reported topological information of a mutant protein lacking its original activity, the study was excluded from the evaluation process.

We deliberately did not include proteins with topology determined solely by crystallography (3D structures), since these proteins are covered by databases such as PDB_TM (Tusnady and Simon, 2001). However, there

*To whom correspondence should be addressed.

are a large number of 3D structures in PDB that correspond to non-transmembrane fragments of transmembrane proteins. This information was included in ExTopoDB as additional topological information derived from structure. The exact localization of soluble segments was obtained from the articles where the structure was published. The total number of referenced articles was raised this way to 1833. We also include annotation for signal peptides, which is obtained either from Uniprot's annotation (Bairoch *et al*., 2005) of the protein (when the signal peptide is experimentally verified) or from a consensus of two of the top-scoring available predictors, SignalP (Bendtsen *et al*., 2004) and Phobius (Kall *et al*., 2004).

The ExTopoDB web application (freely accessible at http://bioinformatics.biol.uoa.gr/ExTopoDB) is based on the combination of two layers: the underlying level is a Mysql database system which contains all protein data and the upper layer is an Apache-PHP applications server that receives user queries and fetches populated HTML data to the web browser client. Protein information can be accessed through four different entry view formats (FASTA, TEXT, TOPOLOGY, XML) apart from the standard view. Detailed description and annotation of each ExTopoDB entry can be found on the manual page of ExTopoDB.

In order to examine how information present in ExTopoDB may help to calculate more accurate topological models we compared constrained (using the information available in the database) and unconstrained predictions for proteins present in ExTopoDB, that additionally, have a known 3D structure deposited in PDB. For topology prediction we used TMHMMfix (Melen *et al*., 2003), HMM-TM (Bagos *et al*., 2006), HMMTOP (Tusnady and Simon, 2001) and PHOBIUS (Kall *et al*., 2004), which allow the incorporation of experimental information as constrains to the prediction procedure. ExTopoDB holds 45 proteins with a PDB structure, which includes the entire protein or a transmembrane fragment of it. Constrained predictions were performed using the available topological information for each protein, derived from the ExTopoDB database. To avoid biased results and perform an independent analysis, we removed proteins that were present in the training set of each prediction algorithm and performed redundancy checks using BLAST (<30%) in order to create four different non-redundant test sets. We came up with 31 proteins for TMHMMfix, 22 proteins for HMMTOP, 26 proteins for HMM-TM and 17 proteins for PHOBIUS (Supplementary Table 1). We have to emphasize that the purpose of this analysis was to evaluate the influence of incorporating topological information present in ExTopoDB to prediction methods in general (since data from ExTopoDB can be used in conjunction with every suitable prediction algorithm) and not to compare the performance of the prediction algorithms used. For the evaluation of the prediction performance in each case we used the Mathew's correlation coefficient (C) and the percentage of correctly predicted residues (Q) (Baldi *et al*., 2000), as well as the segment overlap (SOV) measure (Zemla *et al*., 1999).

3 RESULTS

ExTopoDB contains topological information for 2143 α-helical transmembrane proteins originating from 158 different organisms. The application possesses a user-friendly environment, through which the user may retrieve the necessary information, find available resources and cross-references and perform additional tasks such as running prediction algorithms. In the main page of ExTopoDB, the user may find links for the following tools: Navigation, Text Search, BLAST Search and Download. Through the navigation tool, the user has the ability to browse the database by organism or by the experimental method that was used for the determination of the topology of the protein entry. The user may submit advanced queries for text search and there is an interface for running BLAST against the database (Altschul *et al*., 1997). Each record contains information concerning the protein sequence with cross-references to many publicly available databases. Furthermore, the results of topology prediction using the HMM-TM algorithm (Bagos *et al*., 2006) are included for each protein in the database (unconstrained prediction) and we also incorporated the experimental information about the topology of the proteins, in the HMM-TM prediction procedure, producing more reliable topology models (constrained prediction). There is also an interface that allows user-defined constrained topology prediction using HMM-TM. Each record of ExTopoDB contains the topological information in relation to the experimental technique used and links to the Pubmed entry of each study. The database can be downloaded in several formats through the Download tool.

The incorporation of experimental information in the prediction procedure improved the prediction performance for all measures evaluated, whereas constrained predictions are more reliable (Supplementary Table 2) as previously shown (Melen *et al*., 2003). Moreover, the fraction of correctly predicted residues (Q) increased, for TMHMMfix from 0.838 to 0.858 (2%), for HMM-TM from 0.818 to 0.857 (3.9%), for HMMTOP from 0.788 to 0.861 (7.3%) and for PHOBIUS from 0.874 to 0.884 (1%). The improvement in the prediction performance is more evident in the number of correctly predicted topologies, which perhaps is a metric with increased practical importance. All four prediction methods improve by ∼30% in predicting the correct topology after the incorporation of topology data from ExTopoDB (TMHMMfix 29.1%, HMM-TM 30.8%, HMMTOP 31.8%, PHOBIUS 35.3%). We also observe a 3–9% increase in the SOV measure (Supplementary Table 2), for the different methods used.

4 DISCUSSION

In this article, we present a database, ExTopoDB, which is an up to date collection of experimentally verified topological information for α-helical transmembrane proteins. Compared to similar data sets, ExTopoDB contains the largest number of proteins with topology information derived from 1833 studies. While developing ExTopoDB a similar collection for transmembrane proteins, TOPDB, was published in 2008 by Tusnady and coworkers (Tusnady *et al*., 2008). TOPDB contains records for 1497 transmembrane proteins (1452 α-helical TM proteins and 45 transmembrane β-barrels) with information gathered from the literature and from public databases and has not been updated since 2007. Though TOPDB and ExTopoDB share the same scope, they have only 886 entries in common due to the different criteria used for inclusion. Moreover, even for common entries, ExTopoDB holds a larger number of experiments for each entry, as most information was gathered from the literature (Table 1). Specifically, for the 886 common entries, ExTopoDB collects data from 387 published studies, whereas TOPDB has information from only 181 studies (Supplementary Table 3). Both databases include the study of Daley and coworkers (Daley *et al*., 2005), where C-terminal tagging using alkaline phosphatase (PhoA) and green fluorescent protein (GFP) was used to establish periplasmic or cytoplasmic C-terminal localization for 601 membrane proteins from *Escherichia coli*. ExTopoDB holds data only for proteins where both a PhoA and a GFP clone was available. A large proportion (477 proteins) of the 886 common entries are derived from this study. On top of that, in ExTopoDB there are an additional number of 152 independently identified references for these 477 entries.

Table 1. Entry counts of ExTopoDB according to experimental type and comparison with TOPDB

Experimental type	TOPDB	ExTopoDB
Gene fusion	647	1058
Post-translational modification	31	465
Proteolysis	63	16
Epitope mapping	66	135
Chemical modification	21	50
Structure*	820	520
Other	22	88
Total	1497	2143

Chemical modification experiments refer mostly to cysteine scanning mutagenesis and post-translational modification techniques include glycosylation, phosphorylation and biotinylation analysis. ExTopoDB does not contain transmembrane proteins based solely on the existence of known 3D structure, but only transmembrane proteins of which a non-transmembrane fragment has been crystallized (*).

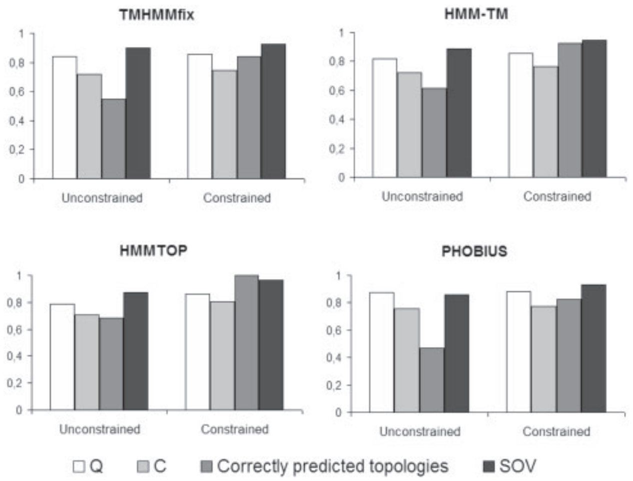

Fig. 1. Measures of accuracy obtained for constrained and unconstrained predictions using TMHMMfix, HMMTOP, HMM-TM and PHOBIUS. Constrained predictions were performed using topology data from ExTopoDB as described in the 'Methods' section.

Our analysis showed that topology data derived from ExTopoDB can be used for generation of constrained predictions, which are more reliable and more accurate at defining the exact locations of the transmembrane segments (Fig. 1). Similar results were also obtained when we used for comparison the data deposited in the MPTOPO database (Jayasinghe *et al.*, 2001), since we observe a high level of agreement in the predicted topology for 31 out of 33 transmembrane proteins with known 3D structure (Supplementary Table S4).

The database might be a valuable tool for researchers, in order to design new experiments (Nilsson *et al.*, 2002) and for bioinformaticians since it provides a large representative set that can be used for training and testing prediction algorithms. For instance, 1056 out of the 2143 proteins have a reliability ≥95% (Melen *et al.*, 2003) and among them, 798 are multispanning ones. Moreover, topology data in ExTopoDB can be used for generating constrained predictions using other methods such as TMHMMfix, HMMTOP and PHOBIUS. Lastly, we have to mention that detailed knowledge of topological information, coupled with prediction algorithms as well as information concerning amino acid contacts derived from biophysical experiments, can be used to calculate structural models that closely resemble the 3D structure (Sorgen *et al.*, 2002). Thus, such data can be of help in future structural studies. We plan to update ExTopoDB once a year.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped in improving the quality of the article.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bagos,P.G. *et al*. (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics*, **7**, 189.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Baldi,P. *et al*. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bendtsen,J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **340**, 783–795.

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

Bernsel,A. *et al*. (2008) Prediction of membrane-protein topology from first principles. *Proc. Natl Acad. Sci. USA*, **105**, 7177–7181.

Daley,D.O. *et al*. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science*, **308**, 1321–1323.

Hessa,T., *et al*. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, **450**, 1026–1030.

Ikeda,M. *et al*. (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res*., **31**, 406–409.

Jayasinghe,S. *et al*. (2001) MPtopo: A database of membrane protein topology. *Protein Sci*., **10**, 455–458.

Kall,L. *et al*. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol*., **338**, 1027–1036.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.

Lee,A.G. (2003) Lipid-protein interactions in biological membranes: a structural perspective. *Biochim. Biophys. Acta*, **1612**, 1–40.

Lomize,M.A. *et al*. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Melen,K. *et al*. (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol*., **327**, 735–744.

Moller,S. *et al*. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Nilsson,J. *et al*. (2002) Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci*., **11**, 2974–2980.

Sorgen,P.L. *et al*. (2002) An approach to membrane protein structure without crystals. *Proc. Natl Acad. Sci. USA*, **99**, 14037–14040.

Tusnady,G.E. *et al*. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*., **33**, D275–278.

Tusnady,G.E. *et al*. (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res*., **36**, D234–239.

Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

van Geest,M. and Lolkema,J.S. (2000) Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiol. Mol. Biol. Rev*., **64**, 13–33.

Zemla,A. *et al*. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.